



SUSE Linux Enterprise Server 15 SP6

ストレージ管理ガイド

ストレージ管理ガイド

SUSE Linux Enterprise Server 15 SP6

このガイドでは、SUSE Linux Enterprise Serverでストレージデバイスを管理する方法について説明します。

発行日: 2025年12月11日

<https://documentation.suse.com>

Copyright © 2006–2025 SUSE LLC and contributors. All rights reserved.

この文書は、GNUフリー文書ライセンスのバージョン1.2または(オプションとして)バージョン1.3の条項に従って、複製、頒布、および/または改変が許可されています。ただし、この著作権表示およびライセンスは変更せずに記載すること。ライセンスバージョン1.2のコピーは、「GNUフリー文書ライセンス」セクションに含まれています。

SUSEの商標については、<https://www.suse.com/company/legal/> を参照してください。サードパーティ各社とその製品の商標は、所有者であるそれぞれの会社に所属します。商標記号(®、™など)は、SUSEおよびその関連会社の商標を示します。アスタリスク(*)は、第三者の商標を示します。

本書のすべての情報は、細心の注意を払って編集されています。しかし、このことは絶対に正確であることを保証するものではありません。SUSE LLC、その関係者、著者、翻訳者のいずれも誤りまたはその結果に対して一切責任を負いかねます。

目次

序文 **xiv**

- 1 利用可能なマニュアル **xiv**
- 2 ドキュメントの改善 **xiv**
- 3 マニュアルの表記規則 **xv**
- 4 サポート **xvii**
 - SUSE Linux Enterprise Serverのサポートステートメント **xviii** • 技術プレビュー **xix**

I ファイルシステムとマウント **1**

1 Linuxファイルシステムの概要 **2**

- 1.1 用語集 **3**
- 1.2 Btrfs **3**
 - 主な特長 **3** • SUSE Linux Enterprise Server上のルートファイルシステム設定 **4** • ReiserFSおよびExtの各ファイルシステムからBtrfsへのマイグレーション **9** • Btrfsの管理 **10** • サブボリュームに対するBtrfsクオータのサポート **10** • Btrfsでのスワッピング **14** • Btrfs send/receive **14** • データ重複排除のサポート **18** • ルートファイルシステムからのサブボリュームの削除 **19**
- 1.3 XFS **20**
 - XFSフォーマット **21**
- 1.4 Ext2 **22**
- 1.5 Ext3 **22**
 - Ext2からの容易で信頼性の高いアップグレード **23** • Ext2ファイルシステムからExt3への変換 **23**
- 1.6 Ext4 **24**
 - 信頼性とパフォーマンス **24** • Ext4ファイルシステムのinodeサイズとinode数 **25** • Ext4へのアップグレード **27**

- 1.7 ReiserFS 29
- 1.8 OpenZFSとZFS 29
- 1.9 tmpfs 29
- 1.10 サポートされている他のファイルシステム 30
- 1.11 ブロックされるファイルシステム 31
- 1.12 Linux環境での大規模ファイルサポート 32
- 1.13 Linuxのカーネルにおけるストレージの制限 33
- 1.14 未使用のファイルシステムブロックの解放 34
 - 定期TRIM 35 • オンラインTRIM 36
- 1.15 ファイルシステムのトラブルシューティング 36
 - Btrfsエラー: デバイスに空き領域がない 36 • Btrfs: デバイス間でデータのバランスを取る 38 • SSDでデフラグメンテーションしない 39
- 1.16 詳細情報 40

2 ファイルシステムのサイズ変更 41

- 2.1 使用例 41
- 2.2 サイズ変更のガイドライン 41
 - サイズ変更をサポートしているファイルシステム 42 • ファイルシステムのサイズの増加 42 • ファイルシステムのサイズの削減 43
- 2.3 Btrfsファイルシステムのサイズの変更 43
- 2.4 XFSファイルシステムのサイズの変更 44
- 2.5 Ext2、Ext3、またはExt4の各ファイルシステムのサイズの変更 45

3 ストレージデバイスのマウント 47

- 3.1 UUIDの理解 47
- 3.2 udevによる永続的なデバイス名 47
- 3.3 ネットワークストレージデバイスのマウント 48

4 ブロックデバイス操作の多層キャッシング 49

- 4.1 一般的な用語 49
- 4.2 キャッシングモード 50
- 4.3 bcache 51
 - 主な特徴 51 • bcacheデバイスのセットアップ 52 • sysfsを使用するbcacheの設定 53
- 4.4 lvmcache 53
 - lvmcacheの構成 54 • キャッシュプールの削除 55

II 論理ボリューム(LVM) 57

5 LVMの設定 58

- 5.1 論理ボリュームマネージャ(LVM)の理解 58
- 5.2 ボリュームグループの作成 60
- 5.3 論理ボリュームの作成 64
 - シンプロビジョニング論理ボリューム 67 • ミラーリングされたボリュームの作成 68
- 5.4 非ルートLVMボリュームグループの自動アクティブ化 69
- 5.5 既存のボリュームグループのサイズ変更 70
- 5.6 論理ボリュームのサイズ変更 71
- 5.7 ボリュームグループまたは論理ボリュームの削除 73
- 5.8 起動時のLVMの無効化 74
- 5.9 LVMコマンドの使用 74
 - コマンドによる論理ボリュームのサイズ変更 78 • LVMキャッシュボリュームの使用 80
- 5.10 LVM2ストレージオブジェクトへのタグ付け 81
 - LVM2タグの使用 81 • LVM2タグの作成要件 82 • コマンドラインでのタグ構文 82 • 設定ファイル構文 83 • クラスタで簡単なアクティベーション制御にタグを使用する 85 • タグを使用して、クラスタ内の好みのホストでアクティブにする 85

6 LVMボリュームスナップショット 89

- 6.1 ボリュームスナップショットの理解 89
- 6.2 LVMによるLinuxスナップショットの作成 91
- 6.3 スナップショットの監視 91
- 6.4 Linuxスナップショットの削除 92
- 6.5 仮想ホスト上の仮想マシンに対するスナップショットの使用 92
- 6.6 スナップショットをソース論理ボリュームとマージして変更を元に戻すか、前の状態にロールバックする 94

III ソフトウェアRAID 97

7 ソフトウェアRAIDの設定 98

- 7.1 RAIDレベルの理解 98
 - RAID 0 98 • RAID 1 99 • RAID 2およびRAID 3 99 • RAID 4 99 • RAID 5 99 • RAID 6 100 • ネストしたコンプレックスRAIDレベル 100
- 7.2 YaSTによるソフトウェアRAID設定 101
 - RAIDの名前 103
- 7.3 AArch64のRAID 5のストライプサイズの設定 104
- 7.4 ソフトウェアRAIDの監視 104
- 7.5 詳細情報 104

8 ルートパーティション用のソフトウェアRAIDの設定 106

- 8.1 ルートパーティション用のソフトウェアRAIDデバイスを使用するための前提条件 106
- 8.2 ルート(/)パーティションにソフトウェアRAIDデバイスを使用するシステムの設定 107

9 ソフトウェアRAID 10デバイスの作成 113

- 9.1 **mdadmによるネストしたRAID 10デバイスの作成 113**
 - mdadmによるネストしたRAID 10 (1+0)デバイスの作成 114 • mdadmによるネストしたRAID 10 (0+1)デバイスの作成 116
- 9.2 **コンプレックスRAID 10の作成 118**
 - コンプレックスRAID 10のデバイスおよびレプリカの数 119 • レイアウト 120 • YaSTパーティションによるコンプレックスRAID 10の作成 122 • mdadmによるコンプレックスRAID 10の作成 125

10 ディグレードRAIDアレイの作成 128

11 mdadmによるソフトウェアRAIDアレイのサイズ変更 130

- 11.1 **ソフトウェアRAIDのサイズの増加 131**
 - コンポーネントパーティションのサイズの増加 132 • RAIDアレイのサイズの増加 133 • ファイルシステムのサイズの増加 134
- 11.2 **ソフトウェアRAIDのサイズの削減 135**
 - ファイルシステムのサイズの削減 135 • RAIDアレイのサイズの削減 135 • コンポーネントパーティションのサイズの削減 137

12 MDソフトウェアRAID用のストレージエンクロージャLEDユーティリティ 139

- 12.1 **ストレージエンクロージャLED監視サービス 140**
- 12.2 **ストレージエンクロージャLED制御アプリケーション 141**
 - パターン名 142 • デバイスのリスト 145 • 例 146
- 12.3 **詳細情報 146**

13 ソフトウェアRAIDのトラブルシューティング 147

- 13.1 **ディスク障害復旧後の回復 147**

IV ネットワークストレージ 149

14 Linux用iSNS 150

- 14.1 **iSNSのしくみ 150**

- 14.2 Linux用iSNSサーバのインストール 152
- 14.3 iSNS検出ドメインの設定 154
 - iSNS検出ドメインの作成 154 • iSCSIノードの検出ドメインへの追加 155
- 14.4 iSNSサービスの開始 157
- 14.5 詳細情報 157

15 IPネットワークの大容量記憶域: iSCSI 158

- 15.1 iSCSI LIOターゲットサーバとiSCSIイニシエータのインストール 159
- 15.2 iSCSI LIOターゲットサーバのセットアップ 160
 - iSCSI LIOターゲットサービスの起動およびファイアウォールの設定 160 • iSCSI LIOターゲットおよびイニシエータのディスクバリに対する認証の設定 161 • ストレージスペースの準備 163 • iSCSI LIOターゲットグループの設定 164 • iSCSI LIOターゲットグループの変更 168 • iSCSI LIOターゲットグループの削除 169
- 15.3 iSCSIイニシエータの設定 169
 - YaSTを使ったiSCSIイニシエータの設定 170 • 手動によるiSCSIイニシエータの設定 173 • iSCSIイニシエータデータベース 174
- 15.4 targetcli-fbを使用したソフトウェアターゲットの設定 175
- 15.5 インストール時のiSCSIディスクの使用 180
- 15.6 iSCSIのトラブルシューティング 180
 - iSCSI LIOターゲットサーバにターゲットLUNをセットアップする際のポータルエラー 181 • iSCSI LIOターゲットが他のコンピュータで表示されない 181 • iSCSIトラフィックのデータパッケージがドロップされる 181 • LVMでiSCSIボリュームを使用する 182 • 設定ファイルが手動に設定されていると、iSCSIターゲットがマウントされる 182
- 15.7 iSCSI LIOターゲットの用語 183
- 15.8 詳細情報 185

16 Fibre Channel Storage over Ethernet Networks: FCoE 186

- 16.1 インストール時におけるFCoEインタフェースの設定 187

- 16.2 FCoEおよびYaSTのFCoEクライアントのインストール 188
- 16.3 YaSTを使用したFCoEサービスの管理 189
- 16.4 コマンドを使用したFCoEの設定 192
- 16.5 FCoE管理ツールを使用したFCoEインスタンスの管理 193
- 16.6 詳細情報 195

17 NVMe-oF 196

- 17.1 概要 196
- 17.2 NVMe-oFホストの設定 196
 - コマンドラインクライアントのインストール 196 • NVMe-oFターゲットの検出 197 • NVMe-oFターゲットへの接続 197 • マルチパス処理 199
- 17.3 NVMe-oFターゲットの設定 199
 - コマンドラインクライアントのインストール 199 • 設定手順 199 • ターゲット設定のバックアップと復元 202
- 17.4 特定のハードウェアの設定 202
 - 概要 202 • Broadcom 202 • Marvell 203
- 17.5 NVMe-oF over TCPからのブート 204
 - システム要件 204 • インストール 204
- 17.6 詳細情報 205

18 デバイスのマルチパスI/Oの管理 206

- 18.1 マルチパスI/Oの理解 206
 - マルチパスの用語 206
- 18.2 ハードウェアサポート 208
 - マルチパスの実装: デバイスマッパーとNVMe 208 • マルチパス処理のストレージアレイ自動検出 208 • 特定のハードウェアハンドラを必要とするストレージアレイ 209
- 18.3 マルチパス処理のプラニング 209
 - 前提条件 210 • マルチパスのインストールタイプ 210 • ディスク管理タスク 211 • ソフトウェアRAIDと複雑なストレージスタック 212 • 高可用性ソリューション 212

- 18.4 マルチパスシステムでのSUSE Linux Enterprise Serverのインストール 212
 - 接続されているマルチパスデバイスを使用しないインストール 213
 - 接続されているマルチパスデバイスによるインストール 213
- 18.5 マルチパスシステムでのSLEの更新 215
- 18.6 マルチパス管理ツール 215
 - デバイスマッパー・マルチパスモジュール 216
 - **multipathd** ディモン 217
 - **multipath** コマンド 219
 - SCSIの永続的な予約と **mpathpersist** 221
- 18.7 マルチパス処理用システムの設定 222
 - マルチパスサービスの有効化、起動、および停止 222
 - マルチパス処理用SANデバイスの準備 224
 - マルチパスデバイスのパーティションと **kpartx** 225
 - **initramfs** の同期状態の維持 225
- 18.8 マルチパス設定 227
 - /etc/multipath.conf** の作成 227
 - **multipath.conf** 構文 227
 - **multipath.conf** セクション 228
 - **multipath.conf** の変更の適用 230
- 18.9 フェールオーバー、キュー格納、およびフェールバックのポリシーの設定 231
 - スタンドアロンサーバでのキュー・ポリシー 234
 - クラスタ化されたサーバでのキュー・ポリシー 234
- 18.10 パスのグループ化と優先度の設定 235
- 18.11 マルチパス処理のためのデバイスの選択 238
 - multipath.conf** の **blacklist** セクション 239
 - **multipath.conf** の **blacklist exceptions** セクション 240
 - デバイスの選択に影響するその他のオプション 240
- 18.12 マルチパスデバイス名とWWID 242
 - WWIDとデバイスの識別 242
 - マルチパスマップのエイリアスの設定 243
 - 自動生成されるユーザフレンドリ名の使用 243
 - マルチパスマップの参照 244
- 18.13 その他のオプション 246
 - 信頼性の低い(「ぎりぎりの」)パスデバイスの処理 248

18.14	ベストプラクティス 250 設定のベストプラクティス 250 • マルチパスI/Oステータスの解釈 250 • マルチパスデバイスでのLVM2の使用 252 • 停止したI/Oの解決 252 • マルチパスデバイスのMD RAID 253 • 新規デバイスのスキャン(再起動なし) 253
18.15	MPIOのトラブルシューティング 254 デバイス選択の問題の理解 254 • デバイス参照の問題の理解 255 • 緊急モードでのトラブルシューティング手順 256 • 技術情報ドキュメント 259

19 NFS共有ファイルシステム 260

19.1	概要 260
19.2	NFSサーバのインストール 261
19.3	NFSサーバの設定 262 YaSTによるファイルシステムのエクスポート 262 • ファイルシステムの手動エクスポート 264 • NFSでのKerberosの使用 267
19.4	クライアントの設定 268 YaSTによるファイルシステムのインポート 268 • ファイルシステムの手動インポート 269 • パラレルNFS(pNFS) 271
19.5	ファイアウォールの背後でNFSサーバとクライアントを操作する 272 NFS 4.x 273 • NFS 3 273
19.6	NFSv4上でのアクセス制御リストの管理 276
19.7	詳細情報 277
19.8	NFSトラブルシューティングのための情報の収集 277 一般的なトラブルシューティング 277 • 高度なNFSデバッグ 279

20 Samba 282

20.1	用語集 282
20.2	Sambaサーバのインストール 284
20.3	Sambaの起動および停止 284

- 20.4 Sambaサーバの設定 284
 - YaSTによるSambaサーバの設定 285 • サーバの手動設定 287
- 20.5 クライアントの設定 292
 - YaSTによるSambaクライアントの設定 292 • クライアント上へのSMB1/CIFS共有のマウント 292
- 20.6 ログインサーバとしてのSamba 293
- 20.7 Active Directoryネットワーク内のSambaサーバ 294
 - realmを使用したActive Directoryの管理 296
- 20.8 詳細トピック 296
 - systemdを使用したCIFSファイルシステムの自動化 297 • Btrfsでの透過的なファイル圧縮 298 • スナップショット 299
- 20.9 詳細情報 306

21 autofsによるオンデマンドマウント 308

- 21.1 インストール 308
- 21.2 設定 308
 - マスタマップファイル 308 • マップファイル 310
- 21.3 操作とデバッグ 311
 - autofsサービスの制御 311 • 自動マウント機能の問題のデバッグ 312
- 21.4 NFS共有の自動マウント 313
- 21.5 詳細トピック 314
 - /net mount point 314 • ワイルドカードを使用したサブディレクトリの自動マウント 314 • CIFSファイルシステムの自動マウント 315

A GNU licenses 316

序文

1 利用可能なマニュアル

オンラインマニュアル

オンラインマニュアルは<https://documentation.suse.com>にあります。さまざまな形式のマニュアルをブラウズまたはダウンロードできます。



注記: 最新のアップデート

最新のアップデートは、通常、英語版マニュアルで入手できます。

SUSE Knowledgebase

問題が発生した場合は、<https://www.suse.com/support/kb/>でオンラインで入手できる技術情報文書(TID)を確認してください。SUSE Knowledgebaseを検索して、お客様のニーズに応じた既知のソリューションを見つけます。

リリースノート

リリースノートは<https://www.suse.com/releasenotes/>を参照してください。

ご使用のシステムで

オフラインで使用するために、リリースノートはシステム上の`/usr/share/doc/release-notes`でも入手できます。個々のパッケージのマニュアルは、`/usr/share/doc/packages`で入手できます。

「マニュアルページ」には、多くのコマンドについても説明されています。説明を表示するには、`man`コマンドに確認したいコマンドの名前を付加して実行してください。システムに`man`コマンドがインストールされていない場合は、`sudo zypper install man`コマンドでインストールします。

2 ドキュメントの改善

このドキュメントに対するフィードバックや貢献を歓迎します。フィードバックを提供するための次のチャネルが利用可能です。

サービス要求およびサポート

ご使用の製品に利用できるサービスとサポートのオプションについては、<https://www.suse.com/support/>を参照してください。

サービス要求を提出するには、SUSE Customer Centerに登録済みのSUSEサブスクリプションが必要です。<https://scc.suse.com/support/requests>に移動して、ログインし、新規作成をクリックします。

バグレポート

<https://bugzilla.suse.com/>から入手できるドキュメントを使用して、問題を報告してください。

このプロセスを容易にするには、このドキュメントのHTMLバージョンの見出しの横にあるReport an issue (問題を報告する)アイコンをクリックしてください。これにより、Bugzillaで適切な製品とカテゴリが事前に選択され、現在のセクションへのリンクが追加されます。バグレポートの入力を直ちに開始できます。

Bugzillaアカウントが必要です。

ドキュメントの編集に貢献

このドキュメントに貢献するには、このドキュメントのHTMLバージョンの見出しの横にあるEdit source document (ソースドキュメントの編集)アイコンをクリックしてください。GitHubのソースコードに移動し、そこからプルリクエストをオーブンできます。GitHubアカウントが必要です。



注記: Edit source document (ソースドキュメントの編集)は英語でのみ利用可能

Edit source document (ソースドキュメントの編集)アイコンは、各ドキュメントの英語版でのみ使用できます。その他の言語では、代わりにReport an issue (問題を報告する)アイコンを使用してください。

このドキュメントに使用されるドキュメント環境に関する詳細については、リポジトリのREADMEを参照してください。

メール

ドキュメントに関するエラーの報告やフィードバックはdoc-team@suse.com宛に送信していただいてもかまいません。ドキュメントのタイトル、製品のバージョン、およびドキュメントの発行日を記載してください。また、関連するセクション番号とタイトル(またはURL)、問題の簡潔な説明も記載してください。

3 マニュアルの表記規則

このマニュアルでは、次の通知と表記規則が使用されています。

- /etc/passwd: ディレクトリ名とファイル名
- PLACEHOLDER: PLACEHOLDERは、実際の値で置き換えられます。
- PATH: 環境変数
- ls、--help: コマンド、オプションおよびパラメータ
- user: ユーザまたはグループの名前
- package_name: ソフトウェアパッケージの名前
- **Alt**、**Alt + F1**: 押すキーまたはキーの組み合わせ。キーはキーボードのように大文字で表示されます。
- ファイル、ファイル、名前を付けて保存: メニュー項目、ボタン
- **AMD/Intel** この説明は、AMD64/Intel 64アーキテクチャにのみ当てはまります。矢印は、テキストブロックの先頭と終わりを示します。□
- **IBM Z, POWER** この説明は、IBM ZおよびPOWERアーキテクチャにのみ当てはまります。矢印は、テキストブロックの先頭と終わりを示します。□
- Chapter 1, 「Example chapter」: このガイドの別の章への相互参照。
- root特権で実行する必要のあるコマンド。これらのコマンドの先頭にsudoコマンドを置いて、特権のないユーザとしてコマンドを実行することもできます。

```
# command
> sudo command
```

- 特権のないユーザでも実行できるコマンド:

```
> command
```

- コマンドは、行末のバックスラッシュ文字(\)で2行または複数行に分割できます。バックスラッシュは、コマンドの呼び出しが行末以降も続くことをシェルに知らせます。

```
> echo a b \
c d
```

- コマンド(プロンプトで始まる)と、シェルによって返される各出力の両方を示すコードブロック:

```
> command
```

- 通知



警告: 警告の通知

続行する前に知っておくべき、無視できない情報。セキュリティ上の問題、データ損失の可能性、ハードウェアの損傷、または物理的な危険について警告します。



重要: 重要な通知

続行する前に知っておくべき重要な情報です。



注記: メモの通知

追加情報。たとえば、ソフトウェアバージョンの違いに関する情報です。



ヒント: ヒントの通知

ガイドラインや実際的なアドバイスなどの役に立つ情報です。

- コンパクトな通知



追加情報。たとえば、ソフトウェアバージョンの違いに関する情報です。



ガイドラインや実際的なアドバイスなどの役に立つ情報です。

4 サポート

SUSE Linux Enterprise Serverのサポートステートメントと、技術レビューに関する一般情報を以下に示します。製品ライフサイクルの詳細については、<https://www.suse.com/lifecycle> を参照してください。

サポート資格をお持ちの場合、<https://documentation.suse.com/sles-15/html/SLES-all/cha-adm-support.html> を参照して、サポートチケットの情報を収集する方法の詳細を確認してください。

4.1 SUSE Linux Enterprise Serverのサポートステートメント

サポートを受けるには、SUSEの適切な購読が必要です。利用可能な特定のサポートサービスを確認するには、<https://www.suse.com/support/>にアクセスして製品を選択してください。

サポートレベルは次のように定義されます。

L1

問題の判別。互換性情報、使用サポート、継続的な保守、情報収集、および利用可能なドキュメントを使用した基本的なトラブルシューティングを提供するように設計されたテクニカルサポートを意味します。

L2

問題の切り分け。データの分析、お客様の問題の再現、問題領域の特定、レベル1で解決できない問題の解決、またはレベル3の準備を行うように設計されたテクニカルサポートを意味します。

L3

問題解決。レベル2サポートで特定された製品の欠陥を解決するようにエンジニアリングに依頼して問題を解決するように設計されたテクニカルサポートを意味します。

契約されているお客様およびパートナーの場合、SUSE Linux Enterprise Serverでは、次のものを除くすべてのパッケージに対してL3サポートを提供します。

- 技術レビュー。
- サウンド、グラフィック、フォント、およびアートワーク。
- 追加の顧客契約が必要なパッケージ。
- モジュール「**Workstation Extension**」の一部として出荷される一部のパッケージは、L2サポートのみです。
- 名前が-develで終わるパッケージ(ヘッダファイルや開発者用のリソースを含む)に対しては、メインのパッケージとともにサポートが提供されます。

SUSEは、元のパッケージの使用のみをサポートします。つまり、変更も、再コンパイルもされないパッケージをサポートします。

4.2 技術プレビュー

技術プレビューとは、今後のイノベーションを垣間見ていただくための、SUSEによって提供されるパッケージ、スタック、または機能を意味します。技術プレビューは、ご利用中の環境で新しい技術をテストする機会を参考までに提供する目的で収録されています。私たちはフィードバックを歓迎しています。技術プレビューをテストする場合は、SUSEの担当者に連絡して、経験や使用例をお知らせください。ご入力いただいた内容は今後の開発のために役立させていただきます。

技術プレビューには、次の制限があります。

- 技術プレビューはまだ開発中です。したがって、機能が不完全であったり、不安定であったり、運用環境での使用には適していないかったりする場合があります。
- 技術プレビューにはサポートが**提供されません**。
- 技術プレビューは、特定のハードウェアアーキテクチャでしか利用できないことがあります。
- 技術プレビューの詳細および機能は、変更される場合があります。その結果、技術プレビューのその後のリリースへのアップグレードは不可能になり、再インストールが必要な場合があります。
- SUSEで、プレビューがお客様や市場のニーズを満たしていない、またはエンタープライズ標準に準拠していないことを発見する場合があります。技術プレビューは製品から予告なく削除される可能性があります。SUSEでは、このようなテクノロジーのサポートされるバージョンを将来的に提供できない場合があります。

ご使用の製品に付属している技術プレビューの概要については、<https://www.suse.com/releasenotes>にあるリリースノートを参照してください。

| ファイルシステムとマウント

- 1 Linuxファイルシステムの概要 2
- 2 ファイルシステムのサイズ変更 41
- 3 ストレージデバイスのマウント 47
- 4 ブロックデバイス操作の多層キャッシング 49

1 Linuxファイルシステムの概要

SUSE Linux Enterprise Serverにはいくつかの異なるファイルシステム(Btrfs、Ext4、Ext3、Ext2、XFSなど)が付属しており、そのいずれかを選択することができます。各ファイルシステムには、それぞれ独自の利点と欠点があります。SUSE Linux Enterprise Serverにおける主要ファイルシステムの機能の対照比較については、https://www.suse.com/releasenotes/x86_64/SUSE-SLES/15-SP3/#file-system-comparison (「Comparison of supported file systems (サポートされるファイルシステムの比較)」)を参照してください。この章では、それらのファイルシステムの機能および利点の概要を説明します。

オペレーティングシステム用のデフォルトファイルシステムはBtrfsであり、他はすべてXFSがデフォルトです。また、Extファイルシステムファミリ、およびOCFS2も引き続きサポートします。デフォルトでは、Btrfsファイルシステムは複数のサブボリュームと共に設定されます。ルートファイルシステムでは、Snapperインフラストラクチャを使用して、スナップショットが自動的に有効になります。Snapperの詳細については、『管理ガイド』、第10章「Snapperを使用したシステムの回復とスナップショット管理」を参照してください。

プロ級のハイパフォーマンスのセットアップには、可用性の高いストレージシステムが必要なことがあります。ハイパフォーマンスのクラスタリングシナリオの要件を満たすため、SUSE Linux Enterprise Serverでは、High AvailabilityアドオンにOCFS2 (Oracle Cluster File System 2)とDRBD (Distributed Replicated Block Device)を組み込んでいます。これらの高度なストレージシステムは、本書では扱いません。詳細については、Administration Guide for SUSE Linux Enterprise High Availability (<https://documentation.suse.com/sle-ha-15/html/SLE-HA-all/book-administration.html>) を参照してください。

ただし、すべてのアプリケーションに最適なファイルシステムは存在しません。各ファイルシステムには特定の利点と欠点があり、それらを考慮する必要があります。最も高度なファイルシステムを選択する場合でも、適切なバックアップ戦略が必要です。

本項で使用される**データの完全性**および**データの一貫性**という用語は、ユーザースペースデータ(ユーザが使用するアプリケーションによりファイルに書き込まれるデータ)の一貫性を指す言葉ではありません。ユーザースペースのデータが一貫しているかどうかは、アプリケーション自体が管理する必要があります。

本項で特に指定のない限り、パーティションおよびファイルシステムの設定または変更に必要なすべての手順は、YaSTパーティショナを使用して実行できます(そうすることをお勧めします)。詳細については、『展開ガイド』、第11章「エキスパートパーティショナ」を参照してください。

1.1 用語集

metadata

ファイルシステムが内包するデータ構造です。これにより、すべてのオンディスクデータが正しく構成され、アクセス可能になります。です。ほとんどすべてのファイルシステムに独自のメタデータ構造があり、それが各ファイルシステムに異なるパフォーマンス特性が存在する理由の1つになっています。メタデータが破損しないよう維持するのは、非常に重要なことです。もし破損した場合、ファイルシステム内にあるすべてのデータがアクセス不能になる可能性があるからです。

inode

サイズ、リンク数、ファイルの内容を実際に格納しているディスクブロックへのポインタ、作成日時、変更日時、アクセス日時など、ファイルに関する各種の情報を含むファイルシステムのデータ構造。

ジャーナル(journal)

ファイルシステムのジャーナルは、ファイルシステムがそのメタデータ内で行う変更を特定のログに記録するオンディスク構造です。ジャーナル機能は、システム起動時にファイルシステム全体をチェックする長時間の検索プロセスが不要なため、ファイルシステムの回復時間を大幅に短縮します。ただし、それはジャーナルが再現できる場合に限定されます。

1.2 Btrfs

Btrfsは、Chris Masonが開発したCOW(コピーオンライト)ファイルシステムです。このシステムは、Ohad Rodehが開発したCOWフレンドリーなBツリーに基づいています。Btrfsは、ロギングスタイルのファイルシステムです。このシステムでは、ロックの変更をジャーナリングする代わりに、それらの変更を新しい場所に書き込んで、リンクインします。新しい変更は、最後の書き込みまで確定されません。

1.2.1 主な特長

Btrfsは、次のような耐障害性、修復、容易な管理機能を提供します。

- 書き込み可能なスナップショット。更新適用後に必要に応じてシステムを容易にロールバックしたり、ファイルをバックアップできます。
- サブボリュームのサポート: Btrfsでは、割り当てられたスペースのプールにデフォルトのサブボリュームが作成されます。Btrfsでは、同じスペースプール内で個々のファイルシステムとして機能する追加サブボリュームを作成できます。サブボリュームの数は、プールに割り当てられたスペースによってのみ制限されます。
- **scrub**を使用したオンラインでのチェックと修復の機能が、Btrfsのコマンドラインツールの一部として利用できます。ツリー構造が正しいことを前提として、データとメタデータの完全性を検証します。マウントしたファイルシステム上で、scrubを定期的に実行することができます。これは、通常の操作中にバックグラウンドプロセスとして実行されます。
- メタデータとユーザデータ用のさまざまなRAIDレベル。
- メタデータとユーザデータ用のさまざまなチェックサム。エラー検出が向上します。
- Linux LVM (Logical Volume Manager)ストレージオブジェクトとの統合。
- SUSE Linux Enterprise Server上でのYaSTパーティションおよびAutoYaSTとの統合。その際、MD (複数デバイス)およびDM (デバイスマッパー)の各ストレージ設定ではBtrfsファイルシステムの作成も行われます。
- 既存のExt2、Ext3、およびExt4ファイルシステムからの、オフラインのマイグレーション。
- /bootのブートローダサポート。Btrfsパーティションからの起動を可能にします。
- マルチボリュームBtrfsは、SUSE Linux Enterprise Server 15 SP6では、RAID0、RAID1、およびRAID10プロファイルでサポートされます。それより高いレベルのRAIDは現時点サポートされませんが、将来のサービスパックでサポートされる可能性があります。
- Btrfsのコマンドを使用して、透過圧縮を設定します。

1.2.2 SUSE Linux Enterprise Server上のルートファイルシステム設定

SUSE Linux Enterprise Serverのルートパーティションは、デフォルトでBtrfsとスナップショットを使用して設定されます。スナップショットを使用すると、更新適用後に必要に応じてシステムを容易にロールバックしたり、ファイルをバックアップしたりできます。ス

ナップショットは、『管理ガイド』、第10章「Snapperを使用したシステムの回復とスナップショット管理」で説明するSUSE Snapperインフラストラクチャを使用して簡単に管理できます。SUSEのSnapperプロジェクトの一般情報については、OpenSUSE.orgにあるSnapper Portal wiki (<http://snapper.io>)を参照してください。

スナップショットを使用してシステムをロールバックする場合、ユーザのホームディレクトリ、WebサーバとFTPサーバのコンテンツ、ログファイルなどのデータがロールバック中に失われたり、上書きされたりしないようにする必要があります。それには、ルートファイルシステムでBtrfsサブボリュームを使用します。サブボリュームは、スナップショットから除外できます。インストール時にYaSTによって提示されるSUSE Linux Enterprise Serverのルートファイルシステムのデフォルト設定には、次のサブボリュームが含まれます。これらがスナップショットから除外される理由を次に示します。

/boot/grub2/i386-pc、/boot/grub2/x86_64-efi、/boot/grub2/powerpc-ieee1275、/boot/grub2/s390x-emu

ブートローダ設定のロールバックはサポートされていません。これらのディレクトリは、アーキテクチャ固有です。最初の2つのディレクトリはAMD64/Intel 64マシン上に存在し、その後の2つのディレクトリはそれぞれIBM POWERとIBM Z上に存在します。

/home

/homeが独立したパーティションに存在していない場合、ロールバック時のデータ損失を避けるために除外されます。

/opt

サードパーティ製品は通常、/optにインストールされます。ロールバック時にこれらのアプリケーションがアンインストールされるのを避けるために除外されます。

/srv

WebおよびFTPサーバ用のデータが含まれています。ロールバック時にデータが失われるのを避けるために除外されます。

/tmp

スナップショットから除外される一時ファイルとキャッシュを含むすべてのディレクトリ。

/usr/local

このディレクトリは、ソフトウェアの手動インストール時に使用します。ロールバック時にこれらのインストール済みソフトウェアがアンインストールされるのを避けるために除外されます。

/var

このディレクトリには、ログ、一時キャッシュ、[/var/opt](#)のサードパーティ製品など、多くのバリアブルファイルが含まれており、仮想マシンのイメージとデータベースのデフォルトの場所です。したがって、このサブボリュームはスナップショットからすべてのこのバリアブルデータを除外するように作成され、コピーオンラインが無効になっています。



警告: ロールバックのサポート

SUSEがロールバックをサポートするのは、事前設定されているサブボリュームがまったく削除されていない場合のみです。ただし、YaSTパーティションを使用して、サブボリュームを追加することはできます。

1.2.2.1 圧縮されたBtrfsファイルシステムのマウント

Btrfsファイルシステムは透過的な圧縮をサポートしています。有効にすると、Btrfsは書き込み時にファイルデータを圧縮し、読み込み時にファイルデータを解凍します。

[compress](#)または[compress-force](#)マウントオプションを使用し、圧縮アルゴリズムを[zstd](#)、[lzo](#)、または [zlib](#) (デフォルト)から選択します。 zlib圧縮は、より圧縮率が高く、一方lzo圧縮はより高速でCPU負荷が低くなります。 zstdアルゴリズムは、パフォーマンスはlzoに近く、圧縮率はzlibと同様という最新の妥協案を提供します。

例:

```
# mount -o compress=zstd /dev/sdx /mnt
```

ファイルを作成し、そのファイルに書き込む場合で、圧縮された結果のサイズが未圧縮サイズよりも大きいか等しい場合、Btrfsはこのファイルに以後も書き込みができるように圧縮をスキップします。この動作が必要ない場合、[compress-force](#)オプションを使用します。最初の圧縮できないデータを含むファイルには有効です。

圧縮は、新規ファイルのみに効果があることに注意してください。圧縮なしで書き込まれたファイルは、ファイルシステムが[compress](#)オプションまたは[compress-force](#)オプションを使用してマウントされたときに圧縮されません。また、[nodatacow](#)属性を持つファイルのエクステントは圧縮されません。

```
# chattr +C FILE
# mount -o nodatacow /dev/sdx /mnt
```

暗号化は、圧縮処理とは関係のない独立した処理です。このパーティションにデータを書き込んだら、詳細を印刷してください。

```
# btrfs filesystem show /mnt
btrfs filesystem show /mnt
Label: 'Test-Btrfs'  uuid: 62f0c378-e93e-4aa1-9532-93c6b780749d
      Total devices 1 FS bytes used 3.22MiB
      devid     1 size 2.00GiB used 240.62MiB path /dev/sdb1
```

永続的に設定したい場合、compressオプションまたはcompress-forceオプションを/etc/fstab設定ファイルに追加します。例:

```
UUID=1a2b3c4d /home btrfs subvol=@/home,compress 0 0
```

1.2.2.2 サブボリュームのマウント

SUSE Linux Enterprise Server上のスナップショットからシステムをロールバックするには、まずスナップショットからブートします。これにより、ロールバックを実行する前に、スナップショットを実行しながらチェックできます。スナップショットからブートできるようにするには、サブボリュームをマウントします(通常は不要な操作です)。

[1.2.2項 「SUSE Linux Enterprise Server上のルートファイルシステム設定」](#)の一覧に示されているサブボリューム以外に、@という名前のボリュームが存在します。これは、ルートパーティション(/)としてマウントされるデフォルトサブボリュームです。それ以外のサブボリュームは、このボリュームにマウントされます。

スナップショットからブートすると、@サブボリュームではなく、スナップショットが使用されます。スナップショットに含まれるファイルシステムの部分は、/として読み込み専用でマウントされます。それ以外のサブボリュームは、スナップショットに書き込み可能でマウントされます。この状態は、デフォルトでは一時的なものです。次回の再起動により、前の設定が復元されます。これを永久的なものにするには、snapper rollbackコマンドを実行します。これにより、今回のブートに使用したスナップショットが新しいデフォルトのサブボリュームになり、再起動後はこのサブボリュームが使用されます。

1.2.2.3 空き領域の確認

通常、ファイルシステムの使用量はdfコマンドで確認します。Btrfsファイルシステムでは、dfの出力は誤解を招く可能性があります。生データが割り当てる領域とは別に、Btrfsファイルシステムもメタデータ用の領域を割り当てて使用するからです。

その結果、まだ大量の領域を使用できるように見えても、Btrfsファイルシステムによって領域不足がレポートされることがあります。その場合、メタデータ用に割り当てられた領域はすべて使用されています。Btrfsファイルシステム上の使用済みの領域と使用可能な領域を確認するには、次のコマンドを使用します。

btrfs filesystem show

```
> sudo btrfs filesystem show /
Label: 'ROOT'  uuid: 52011c5e-5711-42d8-8c50-718a005ec4b3
      Total devices 1  FS bytes used 10.02GiB
      devid     1 size 20.02GiB used 13.78GiB path /dev/sda3
```

ファイルシステムの合計サイズとその使用量を表示します。最後の行のこれら2つの値が一致する場合、ファイルシステム上の領域はすべて割り当て済みです。

btrfs filesystem df

```
> sudo btrfs filesystem df /
Data, single: total=13.00GiB, used=9.61GiB
System, single: total=32.00MiB, used=16.00KiB
Metadata, single: total=768.00MiB, used=421.36MiB
GlobalReserve, single: total=144.00MiB, used=0.00B
```

ファイルシステムの割り当て済みの領域(total)および使用済みの領域の値を表示します。メタデータのtotalおよびusedの値がほぼ等しい場合、メタデータ用の領域はすべて割り当て済みです。

btrfs filesystem usage

```
> sudo btrfs filesystem usage /
Overall:
  Device size:          20.02GiB
  Device allocated:    13.78GiB
  Device unallocated:   6.24GiB
  Device missing:        0.00B
  Used:                 10.02GiB
  Free (estimated):    9.63GiB    (min: 9.63GiB)
  Data ratio:            1.00
  Metadata ratio:       1.00
  Global reserve:       144.00MiB    (used: 0.00B)

          Data   Metadata   System
  Id Path    single    single    single  Unallocated
  --  --
  1  /dev/sda3 13.00GiB 768.00MiB 32.00MiB    6.24GiB
  --  --
  Total      13.00GiB 768.00MiB 32.00MiB    6.24GiB
  Used       9.61GiB 421.36MiB 16.00KiB
```

前の2つのコマンドを組み合わせたのと同様のデータを表示します。

詳細については**man 8 btrfs-filesystem**および<https://btrfs.wiki.kernel.org/index.php/FAQ>を参照してください。

1.2.3 ReiserFSおよびExtの各ファイルシステムからBtrfsへのマイグレーション

btrfs-convertツールを使用して、既存のReiserFSまたはExt (Ext2、Ext3、またはExt4)からBtrfsファイルシステムにデータボリュームをマイグレートすることができます。これにより、アンマウントされた(オフライン)ファイルシステムのインプレース変換を実行できます。これにはbtrfs-convertツールとともにブート可能なインストールメディアが必要な場合があります。このツールは元のファイルシステムの空き領域内にBtrfsファイルシステムを構築し、それに含まれているデータに直接リンクします。メタデータを作成するにはデバイスに十分な空き領域が必要です。さもないと変換に失敗します。元のファイルシステムはそのままとなり、Btrfsファイルシステムによって空き領域が占有されることはありません。必要なスペースの量はファイルシステムのコンテンツによって決まりますが、そこに含まれるファイルシステムオブジェクト(ファイル、ディレクトリ、拡張属性)の数によって左右される場合があります。データは直接参照されるため、ファイルシステム上のデータ量は変換に必要なスペースに影響を与えません。ただし、テールパッキングを使用するファイルや2KiBを超えるサイズのファイルは除きます。



警告: ルートファイルシステムの変換は未サポート

ルートファイルシステムをBtrfsに変換することはサポートされておらず、推奨されません。このような変換を自動化することは、さまざまな手順を特定のセットアップに合わせて調整する必要があるため不可能です。このプロセスには、正しいロールバックを行うための複雑な設定が必要であり、/bootがルートファイルシステム上にある必要があります。システムには特定のサブボリュームが必要であるなどの要件があります。そのため、既存のファイルシステムを保持するか、システム全体を最初から再インストールしてください。

元のファイルシステムをBtrfsファイルシステムに変換するには、次のコマンドを実行します。

```
# btrfs-convert /path/to/device
```



重要: チェック /etc/fstab

変換後は、/etc/fstabに記載されている元のファイルシステムへのすべての参照で、デバイスにBtrfsファイルシステムがあることが示されるように調整されていることを確認する必要があります。

変換時には、Btrfsファイルシステムのコンテンツにソースファイルシステムのコンテンツが反映されます。ソースファイルシステムは、`fs_root/reiserfs_saved/image`で作成された関連する読み込み専用イメージを削除するまで保持されます。イメージファイルの実態は、変換前におけるReiserFSファイルシステムの「スナップショット」であり、Btrfsファイルシステムが変更されても変わりません。イメージファイルを削除するには、`reiserfs_saved`サブボリュームを削除します。

```
# btrfs subvolume delete fs_root/reiserfs_saved
```

ファイルシステムを元に戻すには、次のコマンドを使用します。

```
# btrfs-convert -r /path/to/device
```



警告: 失われる変更

Btrfsファイルシステムとしてマウントされているファイルシステムへの変更はすべて失われます。マウント中には負荷分散操作を実行しないでください。さもないと、ファイルシステムが正しく復元されなくなります。

1.2.4 Btrfsの管理

Btrfsは、YaSTパーティショナおよびAutoYaST内に統合されています。これはインストール時に利用可能で、ルートファイルシステム用のソリューションを設定することができます。インストール後に、YaSTパーティショナを使用して、Btrfsのボリュームの参照と管理を行うことができます。

Btrfsの管理ツールは、`btrfsprogs`パッケージ内に用意されています。Btrfsコマンドの使用法については、`man 8 btrfs`、`man 8 btrfsck`、および`man 8 mkfs.btrfs`コマンドを参照してください。Btrfsの機能については、Btrfs wiki (<https://btrfs.wiki.kernel.org>)を参照してください。

1.2.5 サブボリュームに対するBtrfsクオータのサポート

Btrfsルートファイルシステムのサブボリューム(`/var/log`、`/var/crash`、`/var/cache`など)は、通常の操作中に使用可能なすべてのディスクスペースを使用してしまい、システムに不具合を引き起こす可能性があります。この状況を回避するため、SUSE Linux Enterprise ServerではBtrfsサブボリュームに対するクオータのサポートを提供するようになりました。YaSTの提案に従ってルートファイルシステムを設定する場合は、サブボリュームクオータを有効にして、設定する準備が整います。

1.2.5.1 YaSTを使用したBtrfsクォータの設定

YaSTを使用してルートファイルシステムのサブボリュームにクォータを設定するには、次の手順に従います。

1. YaSTを起動し、システム › パーティショナの順に選択し、はいで警告を確認します。
2. 左側のペインで、Btrfsをクリックします。
3. メインウィンドウで、サブボリュームクォータを有効にするデバイスを選択し、下部にある編集をクリックします。
4. btrfsの編集ウィンドウで、サブボリュームのクォータの有効化チェックボックスをオンにして、次へで確定します。



図 1.1: BTRFSクォータの有効化

5. 既存のサブボリュームのリストから、クォータでサイズを制限するサブボリュームをクリックし、下部にある編集をクリックします。
6. Btrfsのサブボリュームの編集ウィンドウで、サイズ制限を有効にし、参照される最大サイズを指定します。受諾をクリックして確認します。

Btrfs sda2 のサブボリュームの編集

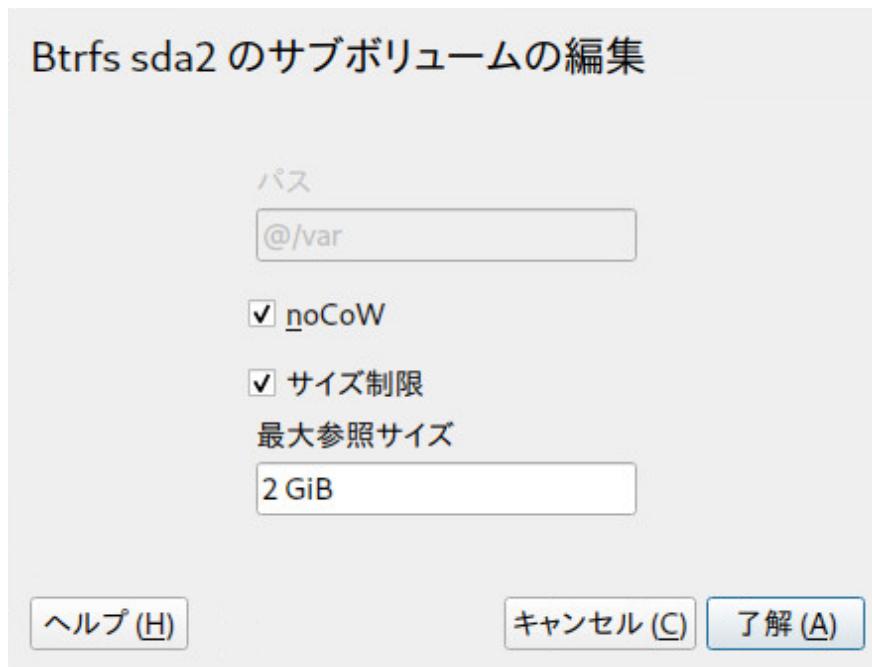


図 1.2: サブボリュームのクォータの設定

サブボリューム名の横に新しいサイズ制限が表示されます。

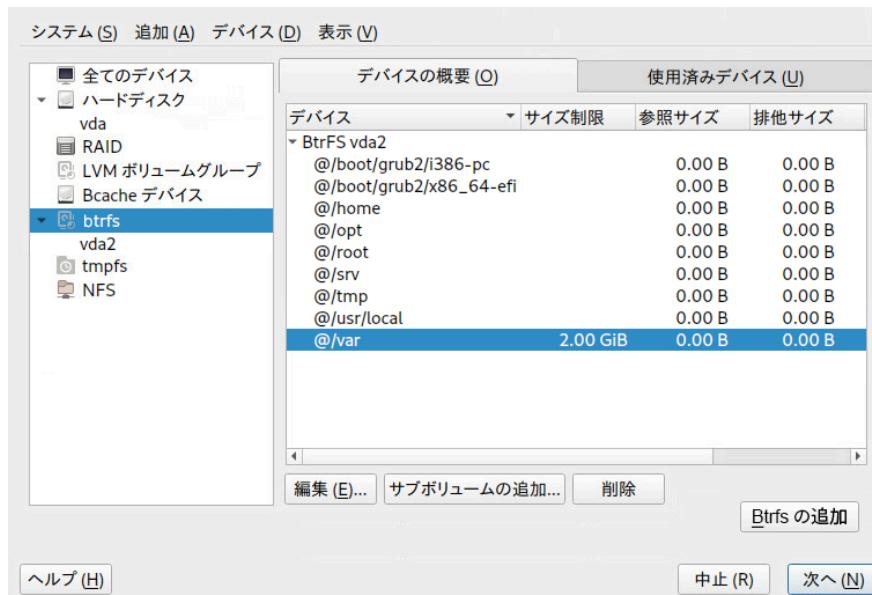


図 1.3: デバイスのサブボリュームのリスト

7. 次へで変更を適用します。

1.2.5.2 コマンドラインでのBtrfsクオータの設定

コマンドラインでルートファイルシステムのサブボリュームにクオータを設定するには、次の手順に従います。

1. クオータサポートを有効にします。

```
> sudo btrfs quota enable /
```

2. サブボリュームのリストを取得します。

```
> sudo btrfs subvolume list /
```

クオータは既存のサブボリュームにのみ設定できます。

3. 前の手順で表示されたサブボリュームの1つにクオータを設定します。サブボリュームは、パス(/var/tmpなど)または0/SUBVOLUME_ID(0/272など)のどちらかによって識別できます。次に、/var/tmpに5 GBのクオータを設定する例を示します。

```
> sudo btrfs qgroup limit 5G /var/tmp
```

サイズは、バイト(50000000000)、キロバイト(5000000K)、メガバイト(5000M)、またはギガバイト(5G)のいずれかの単位で指定できます。結果として得られるサイズは多少異なります。これは、1024バイト=1KiB、1024KiB=1MiBなどだからです。

4. 既存のクオータを一覧にするには、次のコマンドを使用します。max_rfer列に、クオータがバイト単位で表示されます。

```
> sudo btrfs qgroup show -r /
```



ヒント: クオータの無効化

既存のクオータを無効にする場合、クオータサイズをnoneに設定します。

```
> sudo btrfs qgroup limit none /var/tmp
```

特定のパーティションとそのすべてのサブボリュームのクオータサポートを無効にするには、btrfs quota disableを使用します。

```
> sudo btrfs quota disable /
```

1.2.5.3 詳細情報

詳細については、`man 8 btrfs-qgroup`および`man 8 btrfs-quota`を参照してください。Btrfs wiki (UseCases)の<https://btrfs.wiki.kernel.org/index.php/UseCases>ページにも詳細情報が記載されています。

1.2.6 Btrfsでのスワッピング

! 重要: スワッピングを使用したスナップショット

有効なスワップファイルがソースサブボリュームに含まれている場合、スナップショットを作成できません。

SLESでは、結果のスワップファイルに関する次の条件が満たされている場合、Btrfsファイルシステム上のファイルへのスワッピングをサポートしています。

- スワップファイルには、`NODATACOW`および`NODATASUM`マウントオプションが必要です。
- スワップファイルは圧縮できません。`NODATACOW`および`NODATASUM`マウントオプションを設定して、これを確認できます。両方のオプションにより、スワップファイルの圧縮が無効になります。
- スワップファイルは、デバイスのサイズ変更、追加、削除、置換などの排他的な操作の実行中、またはバランシング操作の実行中は有効にできません。
- スワップファイルはスペースにすることはできません。
- スワップファイルはインラインファイルにすることはできません。
- スワップファイルは`single`割り当てプロファイルファイルシステム上にある必要があります。

1.2.7 Btrfs send/receive

Btrfsでは、ファイルシステムの状態をキャプチャするためのスナップショットを作成できます。Snapperでは、たとえばこの機能を使用してシステムの変更前後のスナップショットを作成することで、ロールバックを可能にしています。ただし、send/receive機能とスナップショットを併用すると、リモートの場所にファイルシステムのコピーを作成して管理することもできます。たとえば、この機能を使用してインクリメンタルバックアップを実行できます。

btrfs send操作は、同じサブボリュームの2つの読み込み専用スナップショットの差分を計算して、それをファイルまたはSTDOUTに送信します。**btrfs receive**操作は、sendコマンドの結果を取得して、それをスナップショットに適用します。

1.2.7.1 前提条件

send/receive機能を使用するには、次の要件を満たす必要があります。

- ソース側(send)とターゲット側(receive)にBtrfsファイルシステムが必要です。
- Btrfs send/receiveはスナップショットを操作するため、それぞれのデータがBtrfsサブボリュームに存在する必要があります。
- ソース側のスナップショットは読み込み専用である必要があります。
- SUSE Linux Enterprise 12 SP2以上。それより古いバージョンのSUSE Linux Enterpriseはsend/receiveをサポートしていません。

1.2.7.2 インクリメンタルバックアップ

次の手順では、/data(ソース側)のインクリメンタルバックアップを/backup/data(ターゲット側)に作成する場合を例にして、Btrfs send/receiveの基本的な使用方法を示します。/dataはサブボリュームである必要があります。

手順 1.1: 初期セットアップ

1. ソース側に初期スナップショット(この例ではsnapshot_0という名前)を作成し、それがディスクに書き込まれていることを確認します。

```
> sudo btrfs subvolume snapshot -r /data /data/bkp_data  
sync
```

新しいサブボリューム/data/bkp_dataが作成されます。これは次のインクリメンタルバックアップの基として使用されるので、参照用に保持しておく必要があります。

2. 初期スナップショットをターゲット側に送信します。これは初期のsend/receive操作であるため、完全なスナップショットを送信する必要があります。

```
> sudo bash -c 'btrfs send /data/bkp_data | btrfs receive /backup'
```

ターゲット側に新しいサブボリューム/backup/bkp_dataが作成されます。

初期セットアップが完了したら、インクリメンタルバックアップを作成して、現在のスナップショットと以前のスナップショットの差分をターゲット側に送信できます。手順は常に同じです。

1. ソース側に新しいスナップショットを作成します。
2. 差分をターゲット側に送信します。
3. オプション: 両側のスナップショットの名前変更またはクリーンアップ、あるいはその両方を行います。

手順 1.2: インクリメンタルバックアップの実行

1. ソース側に新しいスナップショットを作成し、それがディスクに書き込まれていることを確認します。次の例では、スナップショットに`bkp_data_CURRENT_DATE`という名前が付いています。

```
> sudo btrfs subvolume snapshot -r /data /data/bkp_data_$(date +%F)  
sync
```

新しいサブボリューム(たとえば、`/data/bkp_data_2016-07-07`)が作成されます。

2. 以前のスナップショットと新たに作成したスナップショットの差分をターゲット側に送信します。そのためには、オプション`-p SNAPSHOT`を使用して、以前のスナップショットを指定します。

```
> sudo bash -c 'btrfs send -p /data/bkp_data /data/bkp_data_2016-07-07 \  
| btrfs receive /backup'
```

新しいサブボリューム`/backup/bkp_data_2016-07-07`が作成されます。

3. その結果、それぞれの側に2つずつ、合計4つのスナップショットが存在することになります。

/data/bkp_data
/data/bkp_data_2016-07-07
/backup/bkp_data
/backup/bkp_data_2016-07-07

続行するには、次の3つのオプションがあります。

- 両方の側のすべてのスナップショットを保持する。このオプションの場合、両方の側のどのスナップショットにもロールバックすることが可能であると同時に、すべてのデータの複製を保持していることになります。これ以上のアクションは必

要ありません。次のインクリメンタルバックアップを実行するときには、最後から2番目のスナップショットをsend操作の親として使用することに注意してください。

- ソース側には最新のスナップショットのみを保持し、ターゲット側にはすべてのスナップショットを保持する。この場合も、両方の側のどのスナップショットにもロールバックできます。ソース側で特定のスナップショットへのロールバックを実行するには、ターゲット側からソース側に、完全なスナップショットのsend/receive操作を実行します。ソース側で削除/移動操作を実行します。
- 両方の側に最新のスナップショットのみを保持する。この方法では、ソース側で作成された最新のスナップショットと同じ状態のバックアップがターゲット側にあります。ほかのスナップショットにロールバックすることはできません。ソース側とターゲット側で削除/移動操作を実行します。
 - a. ソース側に最新のスナップショットのみを保持するには、次のコマンドを実行します。

```
> sudo btrfs subvolume delete /data/bkp_data  
> sudo mv /data/bkp_data_2016-07-07 /data/bkp_data
```

最初のコマンドで以前のスナップショットを削除し、2番目のコマンドで現在のスナップショットの名前を`/data/bkp_data`に変更します。これにより、バックアップされた最新のスナップショットは常に`/data/bkp_data`という名前になります。その結果、常にこのサブボリューム名をインクリメンタルsend操作の親として使用できます。

- b. ターゲット側に最新のスナップショットのみを保持するには、次のコマンドを実行します。

```
> sudo btrfs subvolume delete /backup/bkp_data  
> sudo mv /backup/bkp_data_2016-07-07 /backup/bkp_data
```

最初のコマンドで以前のバックアップスナップショットを削除し、2番目のコマンドで現在のスナップショットの名前を`/backup/bkp_data`に変更します。これにより、最新のバックアップスナップショットは常に`/backup/bkp_data`という名前になります。



ヒント: リモートターゲット側への送信

スナップショットをリモートマシンに送信するには、SSHを使用します。

```
> btrfs send /data/bkp_data | ssh root@jupiter.example.com 'btrfs receive /backup'
```

1.2.8 データ重複排除のサポート

Btrfsはデータ重複排除をサポートします。そのための方法として、ファイルシステム内の複数の同一ブロックを、共通ストレージロケーションにある、そのブロックの1つのコピーを指す論理リンクで置き換えます。SUSE Linux Enterprise Serverでは、ファイルシステムをスキャンして同一ブロックをチェックするduperemoveツールを提供しています。Btrfsファイルシステムで使用される場合、これらのブロックを重複排除して、ファイルシステムのスペースを節約することもできます。duperemoveはデフォルトではインストールされません。使用できるようにするには、パッケージduperemoveをインストールします。



注記: 大量のデータセットの重複排除

大量のファイルを重複排除する場合は、--hashfileオプションを使用します。

```
> sudo duperemove --hashfile HASH_FILE file1 file2 file3
```

--hashfileオプションは、すべての指定されたファイルのハッシュをRAMではなくHASH_FILEに保存して、使い果たされるのを防ぎます。HASH_FILEは再利用可能です。ベースラインハッシュファイルを生成した最初の実行後、大量のデータセットへの変更を非常に迅速に重複排除できます。

duperemoveは、ファイルのリストを処理することも、ディレクトリを再帰的にスキャンすることもできます。

```
> sudo duperemove OPTIONS file1 file2 file3  
> sudo duperemove -r OPTIONS directory
```

動作モードには、読み込み専用と重複排除の2つがあります。読み込み専用モードで実行した場合(-dスイッチを指定しない)、指定されたファイルまたはディレクトリをスキャンして重複ブロックをチェックし、出力します。これは、どのファイルシステムでも機能します。

重複排除モードでのduperemoveの実行は、Btrfsファイルシステムでのみサポートされています。指定されたファイルまたはディレクトリをスキャンした後、重複しているブロックは重複排除用に送信されます。

詳細については、man 8 duperemoveを参照してください。

1.2.9 ルートファイルシステムからのサブボリュームの削除

特定の目的のためにルートファイルシステムからデフォルトのBtrfsサブボリュームの1つを削除する必要がある場合があります。それらの1つはサブボリューム、たとえば`@/home`または`@/srv`を別のデバイスのファイルシステムに変換します。次の手順は、Btrfsサブボリュームを削除する方法を示しています。

- 削除する必要のあるサブボリュームを特定します(たとえば、`@/opt`)。ルートパスのサブボリュームIDが常に「5」であることに注意してください。

```
> sudo btrfs subvolume list /
ID 256 gen 30 top level 5 path @
ID 258 gen 887 top level 256 path @/var
ID 259 gen 872 top level 256 path @/usr/local
ID 260 gen 886 top level 256 path @/tmp
ID 261 gen 60 top level 256 path @/srv
ID 262 gen 886 top level 256 path @/root
ID 263 gen 39 top level 256 path @/opt
[...]
```

- ルートパーティションをホストするデバイス名を見つけます:。

```
> sudo btrfs device usage /
/dev/sda1, ID: 1
Device size:          23.00GiB
Device slack:         0.00B
Data,single:          7.01GiB
Metadata,DUP:         1.00GiB
System,DUP:           16.00MiB
Unallocated:          14.98GiB
```

- ルートファイルシステム(ID 5のサブボリューム)を別のマウントポイント(たとえば`/mnt`)上にマウントします。

```
> sudo mount -o subvolid=5 /dev/sda1 /mnt
```

- マウントされたルートファイルシステムから`@/opt`パーティションを削除します。

```
> sudo btrfs subvolume delete /mnt/@/opt
```

- 以前にマウントされたルートファイルシステムをアンマウントします:。

```
> sudo umount /mnt
```

1.3 XFS

本来は、IRIX OS用のファイルシステムを意図してSGIがXFSの開発を開始したのは、1990年代初期です。XFSの開発動機は、ハイパフォーマンスの64ビットジャーナルファイルシステムの作成により、非常に厳しいコンピューティングの課題に対応することでした。XFSは大規模なファイルを操作する点で非常に優れていて、ハイエンドのハードウェアを適切に活用します。XFSは、SUSE Linux Enterprise Serverのデータパーティション用のデフォルトファイルシステムです。

ただし、XFSの主要機能を一見すれば、XFSが、ハイエンドコンピューティングの分野で、他のジャーナリングファイルシステムの強力な競合相手となっている理由がわかります。

高いスケーラビリティ

XFSはアロケーショングループを使用して高いスケーラビリティを実現する

XFSファイルシステムの作成時に、ファイルシステムの基にあるブロックデバイスは、等しいサイズをもつ8つ以上の線形の領域に分割されます。これらを「アロケーショングループ」と呼びます。各アロケーショングループは、独自のinodeと空きディスクスペースを管理します。実用的には、アロケーショングループを、1つのファイルシステムの中にある複数のファイルシステムと見なすこともできます。アロケーショングループは互いに独立しているものではないため、複数のアロケーショングループをカーネルから同時にアドレス指定できるという特徴があります。この機能は、XFSの高いスケーラビリティに大きく貢献しています。独立性の高いアロケーショングループは、性質上、マルチプロセッサシステムのニーズに適しています。

高いパフォーマンス

XFSはディスクスペースの効率的な管理によって高いパフォーマンスを実現する

空きスペースとinodeは、各アロケーショングループ内のB⁺-Treeによって処理されます。B⁺ツリーの採用は、XFSのパフォーマンスとスケーラビリティを大きく向上させています。XFSでは、プロセスを2分割して割り当てを処理する**遅延割り当て**を使用します。保留されているトランザクションはRAMの中に保存され、適切な量のスペースが確保されます。XFSは、この時点では、データを正確にはどこに(ファイルシステムのどのブロックに)格納するか決定ていません。決定可能な最後の瞬間まで、この決定は遅延(先送り)されます。暫定的に使用される一時データは、ディスクに書き込まれません。XFSがデータの保存場所を決定するまでに、その役割を終えているからです。このように、XFSは、書き込みのパフォーマンスを向上させ、ファイルシステムのフラグメントーションを減少させます。遅延アロケーションは、他のファイルシステムより書き込みイベントの頻度を下げる結果をもたらすので、書き込み中にクラッシュが発生した場合、データ損失が深刻になる可能性が高くなります。

事前割り当てによるファイルシステムの断片化の回避

データをファイルシステムに書き込む前に、XFSはファイルが必要とする空きスペースを予約(プリアロケート、事前割り当て)します。したがって、ファイルシステムの断片化は大幅に減少します。ファイルの内容がファイルシステム全体に分散することがないことで、パフォーマンスが向上します。

1.3.1 XFSフォーマット

SUSE Linux Enterprise Serverは、XFSファイルシステムの「オンディスクフォーマット」(v5)をサポートしています。このフォーマットの主な利点には、全XFSメタデータの自動チェックサム、ファイルタイプのサポート、および1つのファイルに対する大量のアクセス制御リストのサポートがあります。

このフォーマットは、SUSE Linux Enterpriseカーネルの3.12より古いバージョン、xfstoolsの3.2.0より古いバージョン、およびSUSE Linux Enterprise 12より前にリリースされたバージョンのGRUB 2ではサポートされていません。

! 重要: V4は非推奨

XFSではV4フォーマットのファイルシステムが非推奨になっています。このファイルシステムフォーマットは次のコマンドで作成されました。

```
mkfs.xfs -m crc=0 DEVICE
```

このフォーマットはSLE 11以前のリリースで使用されていたもので、現在はdmesgによって警告メッセージが生成されます。

```
Deprecated V4 format (crc=0) will not be supported after September 2030
```

dmesgコマンドの出力に上記のメッセージが表示されたら、ファイルシステムをV5フォーマットに更新することをお勧めします。

1. データを別のデバイスにバックアップします。
2. そのデバイスにファイルシステムを作成します。

```
mkfs.xfs -m crc=1 DEVICE
```

3. 更新したデバイスにバックアップからデータを復元します。

1.4 Ext2

Ext2の起源は、Linuxの歴史の初期にさかのぼります。その前身であったExtended File Systemは、1992年4月に実装され、Linux 0.96cに統合されました。Extended File Systemにはさまざまな変更が加えられてきました。そして、Ext2はLinuxファイルシステムとして数年にわたり非常に高い人気を得ています。その後、ジャーナルファイルシステムが作成され、回復時間が非常に短くなったため、Ext2の重要性は低下しました。

Ext2の利点に関する短い要約を読むと、かつて幅広く好まれ、そして今でも一部の分野で多くのLinuxユーザから好まれるLinuxファイルシステムである理由を理解するのに役立ちます。

堅実性と速度

「古くからある標準」であるExt2は、さまざまな改良が加えられ、入念なテストが実施されてきました。だからこそ、Ext2は非常に信頼性が高いとの評価を得ることが多いのでしょう。ファイルシステムが正常にアンマウントできず、システムが機能停止した場合、e2fsckはファイルシステムのデータの分析を開始します。メタデータは一貫した状態に戻り、保留されていたファイルとデータブロックは、指定のディレクトリ(lost +found)に書き込まれます。ジャーナルファイルシステムとは対照的に、e2fsckは、最近変更されたわずかなメタデータだけではなく、ファイルシステム全体を分析します。この結果、ジャーナルファイルシステムがログデータだけをチェックするのに比べて、かなり長い時間を要します。ファイルシステムのサイズにもよりますが、この手順は30分またはそれ以上を要することがあります。したがって、高可用性を必要とするどのようなサーバでも、Ext2を選択することは望ましくありません。ただし、Ext2はジャーナルを維持せず、わずかなメモリを使用するだけなので、他のファイルシステムより高速なことがあります。

容易なアップグレード性

Ext3は、Ext2のコードをベースとし、Ext2のオンディスクフォーマットとメタデータフォーマットも共用するので、Ext2からExt3へのアップグレードは非常に容易です。

1.5 Ext3

Ext3は、Stephen Tweedieによって設計されました。他のすべての次世代ファイルシステムとは異なり、Ext3は完全に新しい設計理念に基づいているわけではありません。Ext3は、Ext2をベースとしています。これら2つのファイルシステムは、非常に似ています。Ext3ファイルシステムを、Ext2ファイルシステムの上に構築することも容易です。Ext2とExt3の最も重要な違いは、Ext3がジャーナルをサポートしていることです。要約すると、Ext3には、次の3つの主要な利点があります。

1.5.1 Ext2からの容易で信頼性の高いアップグレード

Ext2のコードは、Ext3が次世代ファイルシステムであることを明確に主張するための強力な土台になりました。Ext3では、Ext2の信頼性および堅実性がExt3で採用されたジャーナルファイルシステムの利点とうまく統合されています。XFSのような他のジャーナリングファイルシステムへの移行はかなり手間がかかります(ファイルシステム全体のバックアップを作成し、移行先ファイルシステムを新規に作成する必要があります)が、それとは異なり、Ext3への移行は数分で完了します。ファイルシステム全体を新たに作成し直しても、それが完璧に動作するとは限らないので、Ext3への移行は非常に安全でもあります。ジャーナルファイルシステムへのアップグレードを必要とする既存のExt2システムの数を考慮に入れるに、多くのシステム管理者にとってExt3が重要な選択肢となり得る理由が容易にわかります。Ext3からExt2へのダウングレードも、アップグレードと同じほど容易です。Ext3ファイルシステムのアンマウントを正常に行い、Ext2ファイルシステムとして再マウントします。

1.5.2 Ext2ファイルシステムからExt3への変換

Ext2ファイルシステムをExt3に変換するには、次の手順に従います。

1. Ext3ジャーナルの作成には、`tune2fs -j`をrootユーザとして実行します。

この結果、デフォルトのパラメータを使用してExt3ジャーナルが作成されます。

ジャーナルのサイズおよびジャーナルを常駐させるデバイスを指定するには、`tune2fs -J`とともに適切なジャーナルオプション`size=`および`device=`を指定して、実行します。`tune2fs`プログラムの詳細については、`tune2fs`のマニュアルページを参照してください。

2. ファイル`/etc/fstab`をrootユーザとして編集して、該当するパーティションに指定されているファイルシステムタイプを`ext2`から`ext3`に変更し、その変更内容を保存します。

これにより、Ext3ファイルシステムが認識されるようになります。この変更結果は、次の再起動後に有効になります。

3. Ext3パーティションとしてセットアップされたルートファイルシステムをブートするには、`ext3`と`jbd`の各モジュールを`initrd`に追加します。それには、次を実行します。

- a. `/etc/dracut.conf.d/filesystem.conf`を開くか作成し、次の行を追加します(先頭の空白に注意)。

```
force_drivers+=" ext3 jbd"
```

- b. `dracut -f`コマンドを実行します。

4. システムを再起動します。

1.6 Ext4

2006年に、Ext4はExt3の後継として登場しました。これは、拡張ファイルシステムバージョンの最新ファイルシステムです。Ext4はもともと、最大1エクスピバイトのサイズのボリューム、最大16テビバイトのサイズのファイル、および無制限の数のサブディレクトリをサポートすることによって、ストレージサイズを増やすように設計されました。Ext4は、従来の直接および間接ブロックポインタの代わりにエクステントを使用して、ファイルコンテンツをマッピングします。エクステントを使用することにより、ディスクからのデータの保存と取得の両方が向上しています。

同時に、遅延ブロック割り当て、ファイルシステムチェックルーチンの大幅な高速化など、いくつかのパフォーマンス強化も図られています。また、Ext4は、ジャーナルチェックサムのサポートおよびナノ秒単位でのタイムスタンプの提供により、信頼性を高めています。Ext4には、Ext2およびExt3との完全な後方互換性があり、どちらのファイルシステムもExt4としてマウントできます。



注記: Ext4でのExt3の機能

Ext3の機能は、Ext4カーネルモジュールのExt4ドライバによって完全にサポートされます。

1.6.1 信頼性とパフォーマンス

他のジャーナルファイルシステムは、「メタデータのみ」のジャーナルアプローチに従っています。つまり、メタデータは常に一貫した状態に保持されますが、ファイルシステムのデータ自体については、一貫性が自動的に保証されるわけではありません。Ext4は、メタデータとデータの両方に注意するよう設計されています。「注意」の度合いはカスタマイズできます。Ext4を`data=journal`モードでマウントした場合、最大の保護(データの完全性)を実現しますが、メタデータとデータの両方がジャーナル化されるので、システムの動作が遅くなります。別のアプローチは、`data=ordered`モードを使用することです。これは、データとメタデータ両方の完全性を保証しますが、ジャーナルを適用するのはメタデータのみです。ファイルシステムドライバは、1つのメタデータの更新に対応するすべてのデータブロックを収集します。これらのブロックは、メタデータの更新前にディスクに書き込まれます。その結果、パフォーマンスを犠牲にすることなく、メタデータとデータの両方に関する一貫性を達成できます。3番目のマウントオプションは、`data=writeback`を使用することです。これは、対応す

るメタデータをジャーナルにコミットした後で、データをメインファイルシステムに書き込むことを可能にします。多くの場合、このオプションは、パフォーマンスの点で最善と考えられています。しかし、内部のファイルシステムの完全性が維持される一方で、クラッシュと回復を実施した後では、古いデータがファイル内に再登場させてしまう可能性があります。Ext4では、デフォルトとして、data=orderedオプションを使用します。

1.6.2 Ext4ファイルシステムのinodeサイズとinode数

inodeには、ファイルシステム内のファイルとそのブロック位置に関する情報が格納されます。拡張属性とACL用にinode内のスペースを確保するために、デフォルトのinodeサイズは256バイトに拡大されました。

新規のExt4ファイルシステムを作成する際、inodeテーブル内のスペースは、作成可能なinodeの総数に対して事前に割り当てられています。バイト数/inode数の比率と、ファイルシステムのサイズによって、inode数の上限が決まります。ファイルシステムが作成されると、バイト数/inode数のバイト数の各スペースに対して、1つのinodeが作成されます。

```
number of inodes = total size of the file system divided by the number of bytes per inode
```

inodeの数によって、ファイルシステム内に保有できるファイルの数が決まります。つまり、各ファイルにつき1つのinodeです。

! 重要: 既存のExt4ファイルシステムにおけるinodeサイズの変更は不可能

inodeの割り当て後は、inodeサイズやバイト数/inode数の比率の設定を変えることはできません。異なる設定のファイルシステムを再度作成するか、ファイルシステムを拡張しない限り、新規のinodeは設定できません。inodeの最大数を超えると、ファイルをいくつか削除するまで、ファイルシステム上に新規のファイルを作成することはできません。

新規のExt4ファイルシステムを作成する際に、inodeのスペース使用をコントロールするためのinodeサイズとバイト数/inode数の比率、およびファイルシステム上のファイル数の上限を指定することができます。ブロックサイズ、inodeサイズ、およびバイト数/inode数の比率が指定されない場合は、/etc/mkfs.confファイル内のデフォルト値が適用されます。詳細については、mke2fs.conf(5)マニュアルページを参照してください。

次のガイドラインを使用します。

- **inodeサイズ**: デフォルトのinodeサイズは256バイトです。2の累乗で、ブロックサイズ以下の128以上のバイト数の値を指定します(128、256、512など)。Ext4ファイルシステムで拡張属性またはACLを使用しない場合は、128バイトのみを使用してください。
- **バイト数/inode数の比率**: デフォルトのバイト数/inode数の比率は、16384バイトです。有効なバイト数/inode数の比率は、2の累乗で1024バイト以上(1024、2048、4096、8192、16384、32768など)です。この値は、ファイルシステムのブロックサイズより小さくはできません。なぜなら、ブロックサイズは、データを格納するために使用するスペースの最小チャンクだからです。Ext4ファイルシステムのデフォルトのブロックサイズは、4KiBです。

また、格納する必要があるファイルの数とサイズを検討してください。たとえば、ファイルシステムに多数の小さなファイルを持つことになる場合は、バイト数/inode数の比率を小さめに指定すれば、inodeの数を増やすことができます。ファイルシステムに非常に大きなファイルを入れる場合は、バイト数/inode数の比率を大きめに指定できますが、それによって許容されるinodeの数は減ります。

一般的に、inodeの数は、足りなくなるよりは多すぎる方が得策です。inodeの数が少な過ぎてファイルも非常に小さい場合、実際には空であってもディスク上のファイルの最大数に到達してしまいます。inodeの数が多すぎて、ファイルが非常に大きい場合は、空き領域があることが表示されたとしても、それを使うことができません。なぜなら、inode用に確保されたスペースに新規のファイルを作成することはできないからです。

inodeサイズとバイト数/inode数の比率を設定するには、次のいずれかの方法を使用します。

- **すべての新規Ext4ファイルシステムのデフォルト設定を変更する**: テキストエディタで、`/etc/mke2fs.conf`ファイルの`defaults`セクションを変更し、`inode_size`と`inode_ratio`を目的のデフォルト値に設定します。その値が、すべての新規のExt4ファイルシステムに適用されます。例:

```
blocksize = 4096
inode_size = 128
inode_ratio = 8192
```

- **コマンドラインで**: 新しいExt4ファイルシステムを作成する際に、inodeサイズ(`-I 128`)およびバイト数/inode数の比率(`-i 8192`)を、`mkfs.ext4(8)`コマンドまたは`mke2fs(8)`コマンドに渡します。たとえば、次のコマンドのいずれかを使用します。:

```
> sudo mkfs.ext4 -b 4096 -i 8092 -I 128 /dev/sda2
```

```
> sudo mke2fs -t ext4 -b 4096 -i 8192 -I 128 /dev/sda2
```

- **YaSTを使用したインストール時に:** インストール時に新規のExt4ファイルシステムを作成する際に、inodeサイズとバイト数/inode数の比率を渡します。熟練者向けパーティション設定で、パーティションを選択して、編集をクリックします。フォーマットのオプションで、デバイスをフォーマットするExt4を選択し、オプションをクリックします。フォーマットのオプションダイアログで、ブロックサイズ(バイト単位)、inodeごとのバイト数、およびinodeのサイズドロップダウンボックスから、希望の値を選択します。

たとえば、ブロックサイズ(バイト単位)ドロップダウンボックスから4096を選択しinodeごとのバイト数ドロップダウンボックスから8192を選択し、iノードのサイズドロップダウンボックスから128を選択して、OKをクリックします。



1.6.3 Ext4へのアップグレード

! 重要: データのバックアップ

ファイルシステムの更新を実行する前に、ファイルシステムにあるすべてのデータをバックアップします。

手順 1.3: EXT4へのアップグレード

- Ext2またはExt3からアップグレードするには、以下を有効にする必要があります。

EXT4で必要な機能

エクステント

各ファイルを近くに保持して断片化を防ぐために使用されるハードディスク上の連続したブロック

unint_bg

遅延inodeテーブルの初期化

dir_index

大きいディレクトリ用のハッシュされたbツリー検索

Ext2: as_journal

Ext2ファイルシステムでジャーナリングを有効にします。

これらの機能を有効にするには、次のコマンドを実行します。

- Ext3:

```
# tune2fs -O extents,uninit_bg,dir_index DEVICE_NAME
```

- Ext2:

```
# tune2fs -O extents,uninit_bg,dir_index,has_journal DEVICE_NAME
```

- rootによって`/etc/fstab`ファイルが編集されるとき、`ext3`または`ext2`のレコードを`ext4`に変更します。この変更結果は、次の再起動後に有効になります。

- Ext4パーティションにセットアップされたルートファイルシステムをブートするには、`ext4`と`jbd`の各モジュールを`initramfs`に追加します。`/etc/dracut.conf.d/filesystem.conf`を開くか作成し、次の行を追加します。

```
force_drivers+=" ext4 jbd"
```

既存の`dracut initramfs`を上書きする必要があります。そのためには、次のコマンドを実行します。

```
dracut -f
```

- システムを再起動します。

1.7 ReiserFS

ReiserFSのサポートは、SUSE Linux Enterprise Server 15で廃止されました。既存のパーティションをBtrfsにマイグレートするには、[1.2.3項 「ReiserFSおよびExtの各ファイルシステムからBtrfsへのマイグレーション」](#) を参照してください。

1.8 OpenZFSとZFS

OpenZFSファイルシステムもZFSファイルシステムもSUSEではサポートされていません。ZFSはもともとオープンソースライセンス下でSunによってリリースされていましたが、現在のOracle Solaris ZFSはクローズドソースであるため、SUSEでは使用できません。OpenZFS(元のZFSに基づく)はGPLライセンスとの互換性がない CDDLライセンス下にあるため、カーネルに含めることができません。ただし、Btrfsには同様の設計哲学を持つOpenZFSの優れた代替品があり、SUSEによって完全にサポートされています。

1.9 tmpfs

tmpfsは、RAMベースの仮想メモリファイルシステムです。ファイルシステムは一時的なものです。つまり、ファイルはハードディスクに保存されず、ファイルシステムがアンマウントされると、すべてのデータが破棄されます。

このファイルシステムのデータはカーネル内部キャッシュに保存されます。必要なカーネルキャッシュスペースは拡大または縮小できます。

ファイルシステムの特徴は次のとおりです。

- ファイルへのアクセスが非常に速い。
- スワップがtmpfsマウントに対して有効になると、未使用のデータがスワップされる。
- mount -o remount操作中にデータを失うことなく、ファイルシステムサイズを変更できる。ただし、現在の使用量より小さい値にサイズ変更できない。
- tmpfsはTransparent HugePage Support (THP)をサポートする。

詳細については、以下を参照してください。

- kernel documentation (<https://www.kernel.org/doc/html/latest/filesystems/tmpfs.html>) ↗
- man tmpfs

1.10 サポートされている他のファイルシステム

表1.1 「Linux環境でのファイルシステムのタイプ」は、Linuxがサポートしている他のいくつかのファイルシステムを要約したものです。これらは主に、他の種類のメディアや外部オペレーティングシステムとの互換性およびデータの相互交換を保証することを目的としてサポートされています。

表 1.1: LINUX環境でのファイルシステムのタイプ

ファイルシステムのタイプ	説明
<u>iso9660</u>	CD-ROMの標準ファイルシステム。
<u>msdos</u>	<u>fat</u> 、つまり当初はDOSで使用されていたファイルシステムであり、現在はさまざまなオペレーティングシステムで使用されています。
<u>nfs</u>	Network File System (ネットワークファイルシステム)：ネットワーク内の任意のコンピュータにデータを格納でき、ネットワーク経由でアクセスを付与できます。
<u>ntfs</u>	Windows NT file system (NTファイルシステム)：読み取り専用です。
<u>exfat</u>	USBフラッシュドライブやSDカードなど、フラッシュメモリで使用するように最適化されたファイルシステム。
<u>smbfs</u>	Server Message Block (サーバメッセージブロック): Windowsのような製品が、ネットワーク経由でのファイルアクセスを可能にする目的で採用しています。
<u>ufs</u>	BSD、SunOS、およびNextStepで使用されています。読み取り専用モードでサポートされています。
<u>umsdos</u>	UNIX on MS-DOS(MS-DOS上のUNIX) - 標準 <u>fat</u> ファイルシステムに適用され、特別なファイルを作成することによりUNIXの機能(パーティション、リンク、長いファイル名)を実現します。
<u>vfat</u>	Virtual FAT: <u>fat</u> ファイルシステムを拡張したものです(長いファイル名をサポートします)。

1.11 ブロックされるファイルシステム

セキュリティ上の理由によって、自動マウントからブロックされるファイルシステムがあります。これらのファイルシステムは通常、これ以上保持されず、一般的には使用されません。ただし、このファイルシステムのカーネルモジュールをロードできます。これは、カーネル内のAPIの互換性が保たれているためです。ユーザがマウント可能なファイルシステムと、ファイルシステムの自動マウントをリムーバブルデバイス上で組み合わせると、非特権ユーザがカーネルモジュールの自動ロードをトリガし、リムーバブルデバイスに悪意のあるデータが保存されるという状況が発生する可能性があります。

自動マウントが許可されていないファイルシステムのリストを取得するには、次のコマンドを実行します。

```
> sudo rpm -ql suse-module-tools | sed -nE 's/.*/blacklist_fs-(.*).conf/\1/p'
```

mountコマンドを使用してブロックされるファイルシステムでデバイスをマウントしようとすると、このコマンドはエラーメッセージを出力します。次に例を示します。

```
mount: /mnt/mx: unknown filesystem type 'minix' (hint: possibly blacklisted, see
mount(8)).
```

ファイルシステムのマウントを有効にするには、ブロックリストからそのファイルシステムを削除する必要があります。ブロックされるファイルシステムそれぞれに独自の設定ファイルがあります。たとえば、efsでは/lib/modules.d/60-blacklist_fs-efs.confです。ただし、suse-module-toolsパッケージが更新されると、これらのファイルは必ず上書きされるため、編集しないでください。ブロックされるファイルシステムの自動マウントを許可するには、次の方法があります。

- /dev/nullへのシンボリックリンクを作成します。たとえば、efsファイルシステムの場合には次のようにします。

```
> sudo ln -s /dev/null /etc/modules.d/60-blacklist_fs-efs.conf
```

- 設定ファイルを/etc/modprobe.dにコピーします。

```
> sudo cp /lib/modules.d/60-blacklist_fs-efs.conf /etc/modprobe.d/60-blacklist_fs-
efs.conf
```

設定ファイルの次のステートメントをコメント化します。

```
# blacklist omfs
```

ファイルシステムを自動マウントできない場合でも、modprobeを直接使用して、そのファイルシステムの対応するカーネルモジュールをロードできます。

```
> sudo modprobe FILESYSTEM
```

たとえば、cramfsファイルシステムの場合、出力は次のようにになります。

```
unblacklist: loading cramfs file system module
unblacklist: Do you want to un-blacklist cramfs permanently (<y>es/<n>o/n<e>ver)? y
unblacklist: cramfs un-blacklisted by creating /etc/modprobe.d/60-blacklist_fs-cramfs.conf
```

yesを選択すると、modprobeコマンドは、指定したファイルシステムの設定ファイルからシンボリックリンクを作成するスクリプトを/dev/nullに呼び出します。したがって、このファイルシステムはブロックリストから削除されます。

1.12 Linux環境での大規模ファイルサポート

当初、Linuxは、最大ファイルサイズとして2GiB(2^{31} バイト)をサポートしていました。また、ファイルシステムに大規模ファイルサポートが付いていない限り、32ビットシステム上での最大ファイルサイズは2GiBです。

現在、弊社のすべての標準ファイルシステムでは、LFS(大規模ファイルサポート)を提供しています。LFSは、理論的には、 2^{63} バイトの最大ファイルサイズをサポートします。[表1.2「ファイルおよびファイルシステムの最大サイズ\(オンディスクフォーマット、4KiBブロックサイズ\)](#)」では、Linuxのファイルとファイルシステムの、現行のオンディスクフォーマットの制限事項を概説しています。表内の数字は、ファイルシステムで使用しているブロックサイズが、共通規格である4KiBであることを前提としています。異なるブロックサイズを使用すると結果は異なります。スペースブロックを使用している場合、[表1.2「ファイルおよびファイルシステムの最大サイズ\(オンディスクフォーマット、4KiBブロックサイズ\)](#)」に記載の最大ファイルサイズは、ファイルシステムの実際のサイズより大きいことがあります。



注記: バイナリの倍数

このマニュアルでの換算式: 1024バイト = 1KiB、1024KiB = 1MiB、1024MiB = 1GiB、1024GiB = 1TiB、1024TiB = 1PiB、1024PiB = 1EiB(「NIST: Prefixes for Binary Multiples (<https://physics.nist.gov/cuu/Units/binary.html>)」も参照してください)。

表1.2: [ファイルおよびファイルシステムの最大サイズ\(オンディスクフォーマット、4KiBブロックサイズ\)](#)

ファイルシステム(4KiBブロックサイズ)	ファイルシステムの最大サイズ	ファイルの最大サイズ
Btrfs	16EiB	16EiB

ファイルシステム(4KiBブロックサイズ)	ファイルシステムの最大サイズ	ファイルの最大サイズ
Ext3	16TiB	2TiB
Ext4	1EiB	16TiB
OCFS2 (SLE HAで使用可能なクラスタ対応ファイルシステム)	16TiB	1EiB
XFS	16EiB	8EiB
NFSv2 (クライアント側)	8EiB	2GiB
NFSv3/NFSv4 (クライアント側)	8EiB	8EiB

! 重要: 制限

表1.2 「ファイルおよびファイルシステムの最大サイズ(オンディスクフォーマット、4KiBブロックサイズ)」は、ディスクフォーマット時の制限について説明しています。Linuxカーネルは、操作するファイルとファイルシステムのサイズについて、独自の制限を課しています。管理の初期設定には、次のオプションがあります。

ファイルサイズ

32ビットシステムでは、ファイルサイズが2TiB (2^{41} バイト)を超えることはできません。

ファイルシステムのサイズ

ファイルシステムのサイズは、最大 2^{73} バイトまで可能です。しかし、この制限は、現在使用可能なハードウェアが到達可能な範囲を上回っています。

1.13 Linuxのカーネルにおけるストレージの制限

表1.3 「ストレージの制限」に、SUSE Linux Enterprise Serverに関するストレージに関するカーネルの制限をまとめています。

表 1.3: ストレージの制限

ストレージの機能	制限
サポートされるLUNの最大数	ターゲットあたり16384 LUN。
単一LUNあたりのパスの最大数	デフォルトで無制限。それぞれのパスが、通常のLUNとして扱われます。 実際の制限は、ターゲットあたりのLUNの数と、HBAあたりのターゲットの数(ファイバチャネルHBAの場合は16777215)により決まります。
HBAの最大数	無制限。実際の制限は、システムのPCIスロットの量で決まります。
オペレーティングシステムあたりの、デバイスマッパーマルチパス付きパスの最大数(合計)	約1024。実際の数は、各マルチパスデバイスのデバイス番号文字列の長さによって異なります。これはマルチパスツール内のコンパイル時変数であり、この制限が問題となる場合は増やすこともできます。
最大サイズ(ブロックデバイスごと)	最大8EiB。

1.14 未使用のファイルシステムブロックの解放

SSD (Solid-State Drive)およびシンプロビジョニングされたボリュームでは、ファイルシステムによって使用されていないブロックに対してTrimを実行すると効果的です。SUSE Linux Enterprise Serverは、unmapおよびTRIM操作をサポートするすべてのファイルシステムで、これらの操作を完全にサポートします。

一般的に使用される2種類のTRIM (オンラインTRIMと定期TRIM)があります。デバイスをトリムする最も適切な方法は使用例によって異なります。一般的に、定期TRIMの使用をお勧めします(特に、デバイスに十分な未使用ブロックがある場合)。デバイスの容量が一杯に近くなることが多い場合はオンラインTRIMをお勧めします。



重要: デバイスでのTRIMのサポート

TRIMを使用する前に、デバイスがこの操作をサポートしていることを必ず確認してください。確認を怠ると、そのデバイスのデータが喪失する危険性があります。TRIMのサポートを確認するには、次のコマンドを実行します。

```
> sudo lsblk --discard
```

このコマンドを実行すると、使用可能なすべてのブロックデバイスに関する情報が出力されます。DISC-GRANおよびDISC-MAX列の値が0以外の場合、デバイスはTRIM操作をサポートします。

1.14.1 定期TRIM

定期TRIMは、定期的にsystemdによって呼び出されるfstrimコマンドによって処理されます。コマンドを手動で実行することもできます。

定期TRIMをスケジュールするには、次のようにftrim.timerを有効にします。

```
> sudo systemctl enable ftrim.timer
```

systemdは/usr/lib/systemd/systemにユニットファイルを作成します。デフォルトでは、このサービスは1週間に1回実行されます。通常はこれで十分です。ただし、OnCalendarオプションを目的の値に設定することによって頻度を変更できます。

fstrimのデフォルト動作は、ファイルシステムすべてのブロックを破棄することです。コマンドを呼び出すときにオプションを使用してこの動作を変更できます。たとえば、offsetオプションを渡してトリミング手順の開始地点を定義できます。詳細については、man fstrimを参照してください。

fstrimコマンドを実行すると、/etc/fstabファイルに保存されているすべてのデバイスでトリミングを実行できます。これはTRIM操作をサポートしています。この目的でコマンドを呼び出すときには-Aオプションを使用します。

特定のデバイスのトリミングを無効にするには、次のようにX-fstrim.notrimオプションを/etc/fstabファイルに追加します。

```
UID=83df497d-bd6d-48a3-9275-37c0e3c8dc74 / btrfs defaults,X-fstrim.notrim  
0 0
```

1.14.2 オンラインTRIM

デバイスのオンラインTRIMは、デバイスにデータを書き込むたびに実行されます。

デバイスのオンラインTRIMを有効にするには、次のようにdiscardオプションを/etc/fstabファイルに追加します。

```
UID=83df497d-bd6d-48a3-9275-37c0e3c8dc74 / btrfs defaults,discard
```

また、Ext4ファイルシステムでtune2fsコマンドを使用して/etc/fstabにdiscardオプションを設定します。

```
> sudo tune2fs -o discard DEVICE
```

discardオプションを指定してmountによってデバイスがマウントされた場合も、discardオプションは/etc/fstabに追加されます。

```
> sudo mount -o discard DEVICE
```

注記: オンラインTRIMの欠点

discardオプションを使用すると、一部の低品質SSDデバイスの寿命が短くなる場合があります。オンラインTRIMによってデバイスのパフォーマンスに悪影響が及ぶ場合もあります(大量のデータが削除される場合など)。この状況では、消去ブロックが再割り当てされ、その後すぐに、同じ消去ブロックが未使用としても一度マーク付けされる可能性があります。

1.15 ファイルシステムのトラブルシューティング

本項では、ファイルシステムに関するいくつかの既知の問題と、考えられる解決手段について説明します。

1.15.1 Btrfsエラー: デバイスに空き領域がない

Btrfsファイルシステムを使用しているルート(/)パーティションにデータを書き込めなくなります。「No space left on device」というエラーが発生します。

考えられる原因とこの問題の回避策については、この後の各項を参照してください。

1.15.1.1 Snapperスナップショットによるディスク容量の使用

BtrfsファイルシステムでSnapperが動作している場合、「No space left on device」が表示される問題は、通常は、システム上にスナップショットとして保存されているデータが多すぎるために発生します。

Snapperからいくつかのスナップショットを削除することはできますが、スナップショットはすぐには削除されないので、必要な容量が解放されない可能性があります。

Snapperからファイルを削除するには:

1. 端末を開きます。
2. コマンドプロンプトで、たとえば「btrfs filesystem show」と入力します。

```
> sudo btrfs filesystem show
Label: none uuid: 40123456-cb2c-4678-8b3d-d014d1c78c78
Total devices 1 FS bytes used 20.00GB
devid 1 size 20.00GB used 20.00GB path /dev/sda3
```

3. 以下を入力してください。

```
> sudo btrfs fi balance start MOUNTPOINT -dusage=5
```

このコマンドは、データを空またはほぼ空のデータチャンクに再配置して、その容量を回収し、メタデータに再割り当てしようとします。この処理にはしばらくかかります(1TBで数時間)が、処理中もシステムは使用可能です。

4. Snapperのスナップショットを一覧にします。以下を入力してください。

```
> sudo snapper -c root list
```

5. Snapperから1つ以上のスナップショットを削除します。以下を入力してください。

```
> sudo snapper -c root delete SNAPSHOT_NUMBER(S)
```

必ず最も古いスナップショットを最初に削除してください。古いスナップショットほど、多くの容量を使用します。

この問題が発生しないように、Snapperのクリーンアップアルゴリズムを変更できます。詳細については『管理ガイド』、第10章「Snapperを使用したシステムの回復とスナップショット管理」、10.6.1.2項「クリーンアップアルゴリズム」を参照してください。スナップショットクリーンアップを制御する設定値は、EMPTY_*、NUMBER_*、およびTIMELINE_*です。

ファイルシステムディスクでBtrfsとSnapperを使用する場合、標準のストレージ案の2倍のディスク容量を確保しておくことが推奨されます。YaSTパーティショナは、ルートファイルシステムでBtrfsを使用する場合のストレージ案として、自動的に標準の2倍のディスク容量を提案します。

1.15.1.2 ログ、クラッシュ、およびキャッシュのファイルによるディスク容量の使用

システムディスクがデータでいっぱいになりつつある場合は、/var/log、/var/crash、/var/lib/systemd/coredumpおよび/var/cacheからファイルを削除してみてください。

Btrfs rootファイルシステムのサブボリューム(/var/log、/var/crash、および/var/cache)は、通常の操作中に使用可能なすべてのディスクスペースを使用してしまい、システムに不具合を引き起こす可能性があります。この状況を回避するため、SUSE Linux Enterprise Serverではサブボリュームに対するBtrfsクオータのサポートを提供するようになりました。詳細については[1.2.5項 「サブボリュームに対するBtrfsクオータのサポート」](#)を参照してください。

テストおよび開発用のマシンでは、特にアプリケーションが頻繁にクラッシュする場合、コアダンプが保存されている/var/lib/systemd/coredumpを確認することもできます。

1.15.2 Btrfs: デバイス間でデータのバランスを取る

btrfs balanceコマンドは、btrfs-progsパッケージの一部です。次の状況例では、Btrfsファイルシステムのブロックグループのバランスを取ります。

- 600GBがデータで使用される1TBのドライブがあり、さらに1TBのドライブを追加すると仮定します。バランスを取ることで、理論的には、各ドライブに300GBの使用済みスペースができます。
- デバイスには空に近い多数のチャンクがあります。バランスを取ることによりこれらのチャンクがクリアされるまで、それらのスペースは利用できません。
- 使用率に基づいて半分空のブロックグループを圧縮する必要があります。次のコマンドは、使用率が5%以下のブロックグループのバランスを取ります。

```
> sudo btrfs balance start -dusage=5 /
```



ヒント

/usr/lib/systemd/system/btrfs-balance.timer タイマによって、未使用ブロックグループが毎月クリーンアップされます。

- ブロックデバイスのフルでない部分をクリアし、データをより均等に分散する必要があります。
- 異なるRAIDタイプ間でデータを移行する必要があります。たとえば、一連のディスク上のデータをRAID1からRAID5に変換するには、次のコマンドを実行します。

```
> sudo btrfs balance start -dprofiles=raid1,convert=raid5 /
```



ヒント

Btrfsファイルシステム上のデータのバランスを取るデフォルト動作(たとえば、バランスを取る頻度やマウントポイント)を微調整するには、/etc/sysconfig/btrfsmaintenanceを検査してカスタマイズします。関連するオプションは、BTRFS_BALANCE_で開始されます。

btrfs balanceコマンドの使用に関する詳細については、そのマニュアルページ([man 8 btrfs-balance](#))を参照してください。

1.15.3 SSDでデフラグメンテーションしない

Linuxファイルシステムには、データフラグメンテーションを回避するメカニズムがあり、通常はデフラグメントする必要はありません。ただし、データフラグメンテーションを回避できない場合、およびハードディスクのデフラグメンテーションによってパフォーマンスが大幅に向上する場合に使用するケースがあります。

これは従来のハードディスクにのみ適用されます。フラッシュメモリを使用してデータを保存するソリッドステートディスク(SSD)では、ファームウェアによってデータを書き込むチップを判断するアルゴリズムが提供されます。データは通常、ドライブ全体に分散されます。したがって、SSDのデフラグメンテーションは望ましい効果がなく、不要なデータを書き込むことにより、SSDの製品寿命を縮めます。

この理由のため、SUSEではSSDでデフラグメントしないことを明示的にお勧めします。一部のベンダーも、ソリッドステートディスクをデフラグメントすることについて警告しています。これには、次のものが含まれますが、これに限定されません。

- HPE 3PAR StoreServオールフラッシュ
- HPE 3PAR StoreServコンバージドフラッシュ

1.16 詳細情報

ここまでに説明した各ファイルシステムのプロジェクトには、独自のWebページがあります。そこで詳しいドキュメントとFAQ、さらにメーリングリストを参照することができます。

- Kernel.orgのBtrfs Wiki: <https://btrfs.wiki.kernel.org/>
- E2fsprogs: Ext2/3/4 File System Utilities: <https://e2fsprogs.sourceforge.net/>
- OCFS2プロジェクト: <https://oss.oracle.com/projects/ocfs2/>

ファイルシステム(Linuxファイルシステムに限らない)の詳しい比較については、Wikipediaプロジェクトの「Comparison of file systems」 (https://en.wikipedia.org/wiki/Comparison_of_file_systems#Comparison)を参照してください。

2 ファイルシステムのサイズ変更

ファイルシステムのサイズ変更(パーティションまたはボリュームのサイズ変更と混同しないでください)を使用して、物理ボリュームの使用可能な容量を増やしたり、物理ボリュームで増やした使用可能な容量を使用したりできます。

2.1 使用例

パーティションまたは論理ボリュームのサイズ変更には、YaSTパーティショナを使用することをお勧めします。その際、ファイルシステムは自動的にパーティションまたはボリュームの新しいサイズに合わせて調整されます。ただし、YaSTではファイルシステムのサイズ変更はサポートされていないので、次のようなケースでは手動でサイズを変更する必要があります。

- VM Guestの仮想ディスクのサイズを変更した後。
- NAS (Network Attached Storage)のボリュームのサイズを変更した後。
- 手動でパーティションのサイズを変更した後(たとえば、fdiskまたはpartedを使用)、または論理ボリュームのサイズを変更した後(たとえば、lvresizeを使用)。
- Btrfsファイルシステムを縮小する場合(SUSE Linux Enterprise Server 12の時点ではYaSTはBtrfsファイルシステムの拡大のみをサポートしています)。

2.2 サイズ変更のガイドライン

ファイルシステムのサイズ変更には、データを失う可能性をはらむリスクが伴います。



警告: データのバックアップ

データの喪失を避けるには、データを必ずバックアップしてから、サイズ変更タスクを開始します。

ファイルシステムのサイズを変更する場合は、次のガイドラインに従ってください。

2.2.1 サイズ変更をサポートしているファイルシステム

ボリュームに使用可能な容量を増やせるようにするには、ファイルシステムがサイズ変更をサポートしている必要があります。SUSE® Linux Enterprise Serverでは、ファイルシステム Ext2、Ext3、およびExt4に対して、ファイルシステムのサイズ変更ユーティリティを使用できます。このユーティリティは、次のようにサイズの増減をサポートします。

表 2.1: ファイルシステムサイズ変更のサポート

ファイルシステム	ユーティリティ	サイズの増加(拡大)	サイズの削減(縮小)
Btrfs	<u>btrfs filesystem resize</u>	オンライン	オンライン
XFS	<u>xfs_growfs</u>	オンライン	サポートされていません。
Ext2	<u>resize2fs</u>	オンラインまたはオフライン	オフラインのみ
Ext3	<u>resize2fs</u>	オンラインまたはオフライン	オフラインのみ
Ext4	<u>resize2fs</u>	オンラインまたはオフライン	オフラインのみ

2.2.2 ファイルシステムのサイズの増加

デバイス上で使用可能な最大容量までファイルシステムを拡大することも、正確なサイズを指定することもできます。ファイルシステムのサイズを拡大する前に、必ずデバイス、または論理ボリュームのサイズを拡大しておいてください。

ファイルシステムに正確なサイズを指定する場合は、その新しいサイズが次の条件を満たすかどうかを必ず確認してください。

- 新しいサイズは、既存データのサイズより大きくなければなりません。さもないと、データが失われます。
- ファイルシステムのサイズは使用可能な容量より大きくできないので、新しいサイズは、現在のデバイスサイズ以下でなければなりません。

2.2.3 ファイルシステムのサイズの削減

デバイス上のファイルシステムのサイズを削減する際には、新しいサイズが次の条件を満たすかどうかを必ず確認してください。

- 新しいサイズは、既存データのサイズより大きくななければなりません。さもないと、データが失われます。
- ファイルシステムのサイズは使用可能な容量より大きくできないので、新しいサイズは、現在のデバイスサイズ以下でなければなりません。

ファイルシステムが保存されている論理ボリュームのサイズを削減する場合は、デバイス、または論理ボリュームのサイズを削減しようとする前に、必ずファイルシステムのサイズを削減しておきます。

! 重要: XFS

XFSでフォーマットされたファイルシステムのサイズを縮小することはできません。XFSではそのような機能がサポートされていないためです。

2.3 Btrfsファイルシステムのサイズの変更

Btrfsファイルシステムのサイズは、ファイルシステムがマウントされているときに、**btrfs filesystem resize**コマンドを使用して変更できます。ファイルシステムのマウント中にサイズの増加と縮小の両方を実行できます。

- 端末を開きます。
- 変更するファイルシステムがマウントされていることを確認します。
- 次のどちらかの方法で**btrfs filesystem resize**コマンドを使用して、ファイルシステムのサイズを変更します。
 - ファイルシステムのサイズをデバイスの使用可能な最大サイズまで拡張するには、次のように入力します。

```
> sudo btrfs filesystem resize max /mnt
```

- ファイルシステムを特定のサイズに拡張するには、次のコマンドを入力します。

```
> sudo btrfs filesystem resize SIZE /mnt
```

SIZEを目的のサイズ(バイト単位)で置き換えます。50000K(キロバイト)、250M(メガバイト)、2G(ギガバイト)など、値の単位を指定することもできます。または、プラス(+)記号またはマイナス(-)記号を値の前に付けることにより、現在のサイズに対する増減を指定することもできます。

```
> sudo btrfs filesystem resize +SIZE /mnt  
sudo btrfs filesystem resize -SIZE /mnt
```

4. 次のように入力して、マウントされたファイルシステムに対するサイズ変更の効果をチェックします。

```
> df -h
```

ディスクフリー(df)コマンドは、ディスクの合計サイズ、使用されたブロック数、およびファイルシステム上の使用可能なブロック数を表示します。-hオプションは、読みやすい形式でサイズを出力します(1K、234M、2Gなど)。

2.4 XFSファイルシステムのサイズの変更

XFSファイルシステムのサイズは、ファイルシステムがマウントされているときに、xfs_growfsコマンドを使用して増加できます。XFSファイルシステムのサイズを縮小することはできません。

1. 端末を開きます。
2. 変更するファイルシステムがマウントされていることを確認します。
3. xfs_growfsコマンドを使用して、ファイルシステムのサイズを増やします。次に、ファイルシステムのサイズを、利用可能な最大値まで増やす例を示します。その他のオプションについては、man 8 xfs_growfsを参照してください。

```
> sudo xfs_growfs -d /mnt
```

4. 次のように入力して、マウントされたファイルシステムに対するサイズ変更の効果をチェックします。

```
> df -h
```

ディスクフリー(df)コマンドは、ディスクの合計サイズ、使用されたブロック数、およびファイルシステム上の使用可能なブロック数を表示します。-hオプションは、読みやすい形式でサイズを出力します(1K、234M、2Gなど)。

2.5 Ext2、Ext3、またはExt4の各ファイルシステムのサイズの変更

Ext2、Ext3、およびExt4ファイルシステムのサイズは、各パーティションがマウントされているかどうかにかかわらず、`resize2fs`コマンドを使用して増加できます。Extファイルシステムのサイズを減らすには、ファイルシステムをアンマウントする必要があります。

1. 端末を開きます。
2. ファイルシステムのサイズを減らす必要がある場合は、アンマウントします。
3. 次のどちらかの方法で、ファイルシステムのサイズを変更します。
 - ファイルシステムのサイズを`/dev/sda1`と呼ばれるデバイスの、利用可能な最大サイズまで拡大するには、次のように入力します。

```
> sudo resize2fs /dev/sda1
```

`size`パラメータを指定しない場合、サイズはパーティションのサイズにデフォルト設定されます。

- ファイルシステムを特定のサイズに変更するには、次のコマンドを入力します。

```
> sudo resize2fs /dev/sda1 SIZE
```

`SIZE`パラメータは、要求されたファイルシステムの新サイズを指定します。単位を指定しない場合の`size`パラメータの単位は、ファイルシステムのブロックサイズです。オプションとして、`size`パラメータの後に、次の単位指定子の1つを付けることができます。`s`は512バイトのセクタ、`K`はキロバイト(1キロバイトは1024バイト)、`M`はメガバイト、`G`はギガバイトを表します。

サイズ変更が完了するまで待って、続行します。

4. ファイルシステムがマウントされていない場合は、この時点で、ファイルシステムをマウントします。
5. 次のように入力して、マウントされたファイルシステムに対するサイズ変更の効果をチェックします。

```
> df -h
```

ディスクフリー(df)コマンドは、ディスクの合計サイズ、使用されたブロック数、およびファイルシステム上の使用可能なブロック数を表示します。-hオプションは、読みやすい形式でサイズを出力します(1K、234M、2Gなど)。

3 ストレージデバイスのマウント

このセクションでは、デバイスのマウント時に使用されるデバイス識別子の概要、およびネットワークストレージのマウントに関する詳細について説明します。

3.1 UUIDの理解

UUID (Universally Unique Identifier)は、ファイルシステムの128ビットの番号であり、ローカルシステムと他のシステム全体に渡る固有な識別子です。UUIDは、システムハードウェア情報とタイムスタンプをそのシードの一部として、ランダムに生成されます。UUIDは、通常、デバイスに固有なタグを付けるために使用されます。

非永続的な「従来の」デバイス名(`/dev/sda1`など)を使用すると、ストレージを追加したときに、システムがブートできなくなる可能性があります。たとえば、`root (/)`が`/dev/sda1`に割り当てられている場合、SANを接続した後またはシステムにハードディスクを追加した後に、`/dev/sdg1`に再割り当てされる可能性があります。この場合、ブートローダ設定と`/etc/fstab`ファイルを調整する必要があります、そうしないとシステムは起動できなくなります。

デフォルトで、UUIDはブートローダと、ブートデバイスの`/etc/fstab`ファイルで使用されます。UUIDは、ファイルシステムのプロパティであり、ドライブを再フォーマットすれば変更できます。デバイス名のUUIDを使用することの代替案として、IDまたはラベルでデバイスを識別する方法があります。

UUIDは、ソフトウェアRAIDデバイスのアセンブルと起動の基準としても使用できます。RAIDが作成されると、`md`ドライバは、デバイスのUUIDを生成し、その値を`md`スーパー ブロックに保存します。

どのブロックデバイスのUUIDも、`/dev/disk/by-uuid`ディレクトリ内で見つけることができます。たとえば、UUIDエントリは次のようにになります。

```
> ls -og /dev/disk/by-uuid/
lrwxrwxrwx 1 10 Dec  5 07:48 e014e482-1c2d-4d09-84ec-61b3aefde77a -> ../../sda1
```

3.2 udevによる永続的なデバイス名

Linuxカーネル2.6以降、`udev`によって、永続的なデバイス名を使用した動的な`/dev`ディレクトリのユーザースペースソリューションが提供されます。システムに対してデバイスを追加または削除する場合は、ホットプラグシステムの一部として`udev`が実行されます。

ルールのリストが特定デバイス属性との比較に使用されます。udevルールのインフラストラクチャ(/etc/udev/rules.dディレクトリで定義)は、すべてのディスクデバイスに、それらの認識順序や当該デバイスに使用される接続に関わらず、安定した名前を提供します。udevツールは、カーネルが作成するすべての該当ブロックデバイスを調べ、一定のバス、ドライブタイプ、またはファイルシステムに基づいて、ネーミングルールを適用します。udev用の独自ルールを定義する方法については、「Writing udev Rules (https://reactivated.net/writing_udev_rules.html)」を参照してください。

動的なカーネル提供のデバイスノード名に加えて、udevは、/dev/diskディレクトリ内のデバイスをポイントする永続的なシンボリックリンクのクラスを保持します。このディレクトリは、さらに、by-id、by-label、by-path、およびby-uuidの各サブディレクトリに分類されます。



注記: UUIDジェネレータ

udev以外のプログラム(LVMやmdなど)も、UUIDを生成することがあります、それらのUUIDは/dev/diskにリストされません。

udevによるデバイス管理の詳細については、『管理ガイド』、第29章「udevによる動的カーネルデバイス管理」を参照してください。

udevコマンドの詳細については、man 7 udevを参照してください。

3.3 ネットワークストレージデバイスのマウント

ストレージデバイスの種類によっては、systemd.mountがデバイスのマウントを開始する前にネットワークが設定され、利用可能になっている必要があります。これらのタイプのデバイスのマウントを延期するには、該当するそれぞれのネットワークストレージデバイスの/etc/fstabファイルに_netdevオプションを追加します。次に例を示します。

```
mars.example.org:/nfsexport /shared nfs defaults,_netdev 0 0
```

4 ブロックデバイス操作の多層キャッシング

多層キャッシングは、2つ以上の層で構成される複製/分散キャッシングです。1つは低速であるものの安価な回転方式のブロックデバイス(ハードディスク)に代表され、もう1つは高価であるもののデータ操作を高速に実行します(SSDフラッシュディスクなど)。

SUSE Linux Enterprise Serverは、フラッシュデバイスと回転方式のデバイスとの間のキャッシング用に、それぞれ`bcache`および`lvmcache`という2つの異なるソリューションを実装しています。

4.1 一般的な用語

本項では、キャッシング関連機能の説明でよく使用されるいくつかの用語について説明します。

マイグレーション

論理ブロックの主コピーをデバイス間で移動すること。

昇格

低速なデバイスから高速なデバイスへのマイグレーション。

降格

高速なデバイスから低速なデバイスへのマイグレーション。

起点デバイス

大容量で低速なブロックデバイス。古いか、キャッシングデバイス上のコピーとの同期が保たれている(ポリシーによります)、論理ブロックのコピーが常に含まれます。

キャッシングデバイス

小容量で高速なブロックデバイス。

メタデータデバイス

キャッシングに入っているブロック、ダーティブロック、およびポリシーオブジェクトが使用する追加のヒントを記録する小容量のデバイス。この情報はキャッシングデバイスに配置することができますが、別個に保持することにより、ボリュームマネージャで異なった設定にすることができます。たとえば、堅牢性を強化するためのミラーとして設定できます。メタデータデバイスを使用できるキャッシングデバイスは1つだけです。

ダーティブロック

何らかのプロセスがキャッシングに配置されたデータブロックに書き込みを行う場合、そのキャッシングされているブロックは、キャッシング内で上書きされていて、元のデバイスにもう一度書き込む必要があるため、「ダーティ」とマークされます。

キャッシングミス

I/O操作の要求は、まず、キャッシングされたデバイスのキャッシングを参照します。要求された値が見つからなかった場合、デバイス自体を検索しますが、これは低速です。これを「キャッシングミス」と呼びます。

キャッシングヒット

要求された値がキャッシングされたデバイスのキャッシング内で見つかった場合、その値は高速に提供されます。これを「キャッシングヒット」と呼びます。

コールドキャッシング

値が格納されていない(空の)キャッシングのことで、「キャッシングミス」を引き起こします。キャッシングされたブロックデバイスの操作が進むにつれて、キャッシングはデータで満たされていき、「ウォーム」になります。

ウォームキャッシング

すでに何らかの値が格納されていて、「キャッシングヒット」になる確立が高いキャッシング。

4.2 キャッシングモード

多層キャッシングで使用される基本的なキャッシングモードは、「ライトバック」、「ライトスルー」、「ライトアラウンド」、および「パススルー」です。

ライトバック

キャッシングされているブロックに書き込まれたデータは、キャッシングにのみ書き込まれ、そのブロックはダーティとマークされます。これはデフォルトのキャッシングモードです。

ライトスルー

キャッシングされているブロックへの書き込みは、起点デバイスとキャッシングデバイスの両方にヒットするまで完了しません。「ライトスルー」キャッシングでは、クリーンブロックはクリーンな状態のままでです。

ライトアラウンド

ライトスルーキャッシングと同様の手法ですが、書き込みI/Oは、キャッシングをバイパスして永続ストレージに直接書き込まれます。この手法では、直後に再読み込みされない書き込みI/Oによってキャッシングがいっぱいになるのを防ぐことができますが、最近書き込まれたデータの読み込み要求で「キャッシングミス」が発生し、低速なバルクストレージからの読み込みが必要になり、レイテンシが増加するという欠点があります。

パススルー

「パススルー」モードを有効にするには、キャッシングがクリーンである必要があります。読み込みは、キャッシングをバイパスして起点デバイスから実行されます。書き込みは起点デバイスに転送され、キャッシングブロックは「無効化」されます。「パススルー」では、データ整合性が維持されるため、データ整合性を気にすることなくキャッシングデバイスをアクティブ化できます。書き込みが実行されるにつれて、キャッシングは徐々にコールドになります。後でキャッシングの整合性を検証できる場合、または`invalidate_cblocks`メッセージを使用して整合性を保証できる場合は、キャッシングデバイスがまだウォームである間に、デバイスを「ライトスルー」または「ライトバック」モードに切り替えることができます。それ以外の場合は、目的のキャッシングモードに切り替える前に、キャッシングの内容を破棄できます。

4.3 bcache

`bcache`はLinuxカーネルブロック層のキャッシングです。1台以上の高速なディスクドライブ(SSDなど)を1台以上の低速なハードディスクのキャッシングとして動作させることができます。`bcache`は、ライトスルーとライトバックをサポートし、使用するファイルシステムから独立しています。デフォルトでは、SSDの強みである、ランダム読み込みとランダム書き込みのみのキャッシングを実行します。デスクトップやサーバのほか、ハイエンドのストレージアレイにも適しています。

4.3.1 主な特徴

- 1つのキャッシングデバイスを使用して、任意の数のバックキングデバイスをキャッシングできます。バックキングデバイスは、マウント中および使用中のランタイムに接続および切断できます。
- 不正なシャットダウンから回復します。キャッシングがバックキングデバイスと整合性があるようになるまで、書き込みは完了しません。
- 輻輳する場合、SSDへのトラフィックを制限します。
- 非常に効率的なライトバック実装。データは常にソートされた順序で書き込まれます。
- 運用環境での使用における安定性と信頼性。

4.3.2 bcacheデバイスのセットアップ

この項では、bcacheデバイスのセットアップと管理の手順を説明します。

1. bcache-toolsパッケージをインストールします。

```
> sudo zypper in bcache-tools
```

2. バッキングデバイスを作成します(通常は機械式ドライブ)。デバイス全体、パーティション、またはその他の標準ブロックデバイスをバッキングデバイスにすることができます。

```
> sudo make-bcache -B /dev/sdb
```

3. キャッシュデバイスを作成します(通常はSSDディスク)。

```
> sudo make-bcache -C /dev/sdc
```

この例では、デフォルトのブロックサイズとバケットサイズである512Bと128KBを使用しています。ブロックサイズはバッキングデバイスのセクタサイズ(通常は512または4k)と一致している必要があります。バケットサイズは、書き込みの増大を防ぐために、キャッシングデバイスの消去ブロックサイズと一緒にしている必要があります。たとえば、セクタが4kのハードディスクと消去ブロックサイズが2MBのSSDを使用する場合、このコマンドは次のようにになります。

```
sudo make-bcache --block 4k --bucket 2M -C /dev/sdc
```



ヒント: 複数デバイスのサポート

make-bcacheは、複数のバッキングデバイスとキャッシュデバイスを同時に準備および登録できます。この場合、後から手動でキャッシュデバイスをバッキングデバイスに接続する必要はありません。

```
> sudo make-bcache -B /dev/sda /dev/sdb -C /dev/sdc
```

4. bcacheデバイスは次のように表示されます。

```
/dev/bcacheN
```

さらに、次のようにも表示されます。

```
/dev/bcache/by-uuid/UUID  
/dev/bcache/by-label/LABEL
```

bcacheデバイスは通常の方法で正常にフォーマットおよびマウントできます。

```
> sudo mkfs.ext4 /dev/bcache0  
> sudo mount /dev/bcache0 /mnt
```

bcacheデバイスは、/sys/block/bcacheN/bcacheにあるsysfsによって制御できます。

5. キャッシュデバイスとバックингデバイスの両方を登録した後、バックингデバイスを関連キャッシュセットに接続して、キャッシュを有効にする必要があります。

```
> echo CACHE_SET_UUID > /sys/block/bcache0/bcache/attach
```

CACHE_SET_UUIDは/sys/fs/bcacheで確認できます。

6. デフォルトでは、bcacheはパススルーキャッシングモードを使用します。たとえば、これをライトバックに変更するには、次のコマンドを実行します。

```
> echo writeback > /sys/block/bcache0/bcache/cache_mode
```

4.3.3 sysfsを使用するbcacheの設定

bcacheデバイスは、sysfsインターフェースを使用してランタイム設定値を保存します。このようにして、bcacheバックングディスクとキャッシュディスクの動作を変更したり、使用状況の統計を表示したりできます。

bcache sysfsパラメータの全リストについては、/usr/src/linux/Documentation/bcache.txtファイルのコンテンツ、主にSYSFS - BACKING DEVICE、SYSFS - BACKING DEVICE STATS、およびSYSFS - CACHE DEVICEセクションを参照してください。

4.4 lvmcache

lvmcacheは、論理ボリューム(LV)で構成されるキャッシングメカニズムです。dm-cacheカーネルドライバを使用し、ライトスルー(デフォルト)およびライトバックのキャッシングモードをサポートします。lvmcacheは、データの一部をより高速で小容量のLVに動的に移行することによって、大容量で低速なLVのパフォーマンスを向上させます。LVMの詳細については、パートII 「論理ボリューム(LVM)」を参照してください。

LVMでは、この小容量で高速なLVを「キャッシングプールLV」と呼びます。一方、大容量で低速なLVを「起点LV」と呼びます。dm-cacheの要件があるため、LVMは、キャッシングプールLVをさらに「キャッシングデータLV」と「キャッシングメタデータLV」という2つのデバイスに分割し

ます。キャッシュデータLVは、速度の向上を目的として、起点LVからのデータブロックのコピーが保持される場所です。キャッシュメタデータLVには、データブロックが保存されている場所を指定するアカウンティング情報が格納されます。

4.4.1 lvmcacheの構成

この項では、LVMベースのキャッシングの作成と設定の手順を説明します。

1. 起点LVを作成します。新しいLVを作成するか既存のLVを使用して、起点LVにします。

```
> sudo lvcreate -n ORIGIN_LV -L 100G vg /dev/SLOW_DEV
```

2. キャッシュデータLVを作成します。このLVには、起点LVからのデータブロックが格納されます。このLVのサイズがキャッシュのサイズになり、キャッシュプールLVのサイズとして報告されます。

```
> sudo lvcreate -n CACHE_DATA_LV -L 10G vg /dev/FAST
```

3. キャッシュメタデータLVを作成します。このLVには、キャッシュプールメタデータが格納されます。このLVのサイズは、キャッシュデータLVの約1000分の1にする必要があります。最小サイズは8MBです。

```
> sudo lvcreate -n CACHE_METADATA_LV -L 12M vg /dev/FAST
```

これまでに作成したボリュームの一覧を表示します。

```
> sudo lvs -a vg
  LV           VG   Attr      LSize   Pool Origin
  cache_data_lv   vg   -wi-a----  10.00g
  cache_metadata_lv   vg   -wi-a----  12.00m
  origin_lv       vg   -wi-a---- 100.00g
```

4. キャッシュプールLVを作成します。データLVとメタデータLVをキャッシュプールLVに結合します。同時にキャッシュプールLVの動作を設定できます。

CACHE_POOL_LVは、CACHE_DATA_LVの名前を引き継ぎます。

CACHE_DATA_LVはCACHE_DATA_LV_cdataという名前に変更され、非表示になります。

CACHE_META_LVはCACHE_DATA_LV_cmetaという名前に変更され、非表示になります。

```
> sudo lvconvert --type cache-pool \
--poolmetadata vg/cache_metadata_lv vg/cache_data_lv
```

```
> sudo lvs -a vg
  LV           VG   Attr      LSize   Pool Origin
```

```
cache_data_lv      vg  Cwi---C---  10.00g
[cache_data_lv_cdata]  vg  Cwi-----  10.00g
[cache_data_lv_cmeta]  vg  ewi-----  12.00m
origin_lv          vg  -wi-a----- 100.00g
```

5. キャッシュLVを作成します。キャッシュプールLVを起点LVにリンクして、キャッシュLVを作成します。

ユーザがアクセス可能なキャッシュLVは起点LVの名前を引き継ぎ、起点LVは非表示LVになつてORIGIN_LV_corigという名前に変更されます。

キャッシュLVは、ORIGIN_LVの名前を引き継ぎます。

ORIGIN_LVはORIGIN_LV_corigという名前に変更され、非表示になります。

```
> sudo lvconvert --type cache --cachepool vg/cache_data_lv vg/origin_lv
```

```
> sudo lvs -a vg
  LV           VG   Attr       LSize   Pool   Origin
  cache_data_lv      vg  Cwi---C---  10.00g
  [cache_data_lv_cdata]  vg  Cwi-ao---- 10.00g
  [cache_data_lv_cmeta]  vg  ewi-ao---- 12.00m
  origin_lv          vg  Cwi-a-C--- 100.00g  cache_data_lv [origin_lv_corig]
  [origin_lv_corig]    vg  -wi-ao---- 100.00g
```

4.4.2 キャッシュプールの削除

LVキャッシュをオフにする方法はいくつかあります。

4.4.2.1 キャッシュLVからキャッシュプールLVを切断

キャッシュプールLVをキャッシュLVから接続解除して、未使用キャッシュプールLVとキャッシュされていない起点LVを残すことができます。データは、必要に応じてキャッシュプールから起点LVに書き戻されます。

```
> sudo lvconvert --splitcache vg/origin_lv
```

4.4.2.2 起点LVを削除せずにキャッシュプールLVを削除

この方法では、必要に応じてキャッシュプールから起点LVにデータを書き戻してから、キャッシュプールLVを削除し、キャッシュされていない起点LVを残します。

```
> sudo lvremove vg/cache_data_lv
```

次に示す別のコマンドでも、キャッシングLVからキャッシングプールを接続解除し、キャッシングプールを削除します。

```
> sudo lvconvert --uncache vg/origin_lv
```

4.4.2.3 起点LVとキャッシングプールLVの両方を削除

キャッシングLVを削除すると、起点LVとリンクされたキャッシングプールLVの両方が削除されます。

```
> sudo lvremove vg/origin_lv
```

4.4.2.4 詳細情報

サポートされるキャッシングモード、冗長なサブ論理ボリューム、キャッシングポリシー、既存のLVからキャッシングタイプへの変換など、lvmcacheに関連するその他のトピックは、lvmcacheのマニュアルページ([man 7 lvmcache](#))で参照できます。

II 論理ボリューム(LVM)

5 LVMの設定 58

6 LVMボリュームスナップショット 89

5 LVMの設定

この章では、LVM (Logical Volume Manager)の原理と多くの状況で役立つ基本機能を説明します。YaST LVMの設定は、YaST Expert Partitionerからアクセスできます。このパーティショニングツールにより、既存のパーティションを編集、および削除できます。また、LVMで使用する新規パーティションを作成することもできます。



警告: リスク

LVMを使用することでデータ損失などの危険性が増加する恐れがあります。この危険性にはアプリケーションのクラッシュ、電源障害、誤ったコマンドなども含まれます。LVMまたはボリュームの再設定を実施する前にデータを保存してください。バックアップなしでは作業を実行しないでください。

5.1 論理ボリュームマネージャ(LVM)の理解

LVMは、複数の物理ボリューム(ハードディスク、パーティション、LUN)にハードディスクスペースを柔軟に分散することができます。LVMが開発された理由は、インストール中に初期パーティショニングが終了した後でのみ、ハードディスクスペースのセグメンテーションを変更するニーズが発生する可能性があるためです。実行中のシステムでパーティションを変更することは困難なので、LVMは必要に応じて論理ボリューム(LV)を作成できるストレージスペースの仮想プール(ボリュームグループ(VG))を提供します。オペレーティングシステムは物理パーティションの代わりにこれらのLVにアクセスします。ボリュームグループは2つ以上のディスクにまたがることができます。したがって、複数のディスクまたはそれらの一部で1つのVGを構成できます。この方法で、LVMは物理ディスクスペースから一種の抽象化を行います。この抽象化により、物理パーティショニングを使用する場合よりはるかに簡単で安全な方法でセグメンテーションを変更できます。

図5.1 「物理パーティショニング対LVM」では物理パーティショニング(左)とLVM区分(右)を比較しています。左側は、1つのディスクが割り当てられたマウントポイント(MP)をもつ3つの物理パーティション(PART)に分かれています。これによりオペレーティングシステムはそれぞれのパーティションにアクセスできます。右側では2つのディスクがそれぞれ3つの物理パーティションに分かれています。2つのLVMボリュームグループ(VG 1およびVG 2)が定義されています。VG 1にはDISK 1からのパーティションが2つ、DISK 2からのパーティションが1つ含まれています。VG 2には、ディスク2の残りの2つのパーティションが含まれています。

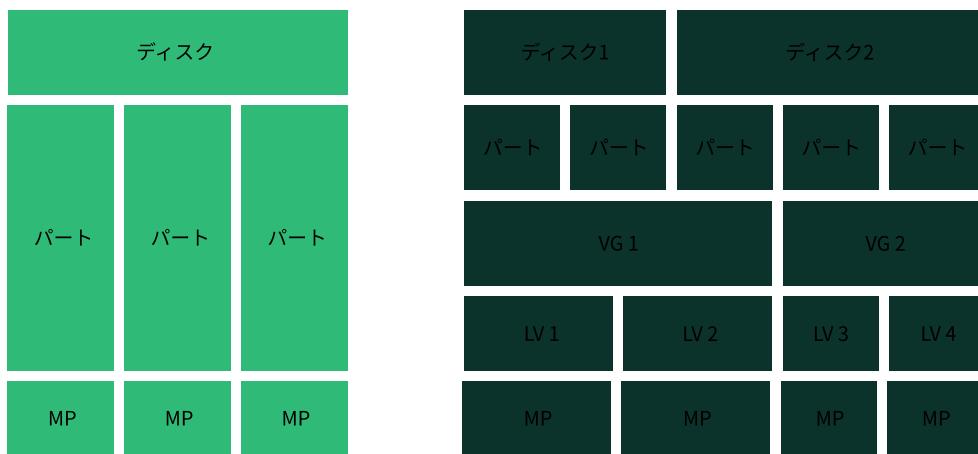


図 5.1: 物理パーティショニング対LVM

LVMでは、ボリュームグループに組み込まれた物理ディスクをPV(物理ボリューム)と呼びます。図5.1「物理パーティショニング対LVM」のボリュームグループ内には、4つの論理ボリューム(LV 1からLV 4)が定義されています。これらのボリュームは、関連付けられたマウントポイント(MP)を介してオペレーティングシステムに使用されます。別の論理ボリュームとの境界とパーティションの境界を並べることはできません。この例ではLV 1およびLV 2の間に境界があります。

LVMの機能:

- 複数のハードディスクまたはパーティションを大きな論理リュームにまとめることができます。
- 提供された設定が適切であれば、LV (/usrなど)は空きスペースがなくなったときに拡張することができます。
- LVMを使用することで、実行中のシステムにハードディスクまたはLVを追加できます。ただし、そのためには、ディスクやLVを追加することのできるホットプラグ可能なハードウェアが必要になります。
- 複数の物理ボリューム上に論理ボリュームのデータストリームを割り当てる**ストライピングモード**を有効にすることもできます。これらの物理ボリュームが別のディスクに存在する場合、RAID 0と同様に読み込みおよび書き込みのパフォーマンスを向上できます。
- スナップショット機能は稼動中のシステムで一貫性のある(特にサーバ)バックアップを取得できます。



注記: LVMとRAID

LVMではRAIDレベル0、1、4、5、6もサポートしていますが、`mdraid`を使用することをお勧めします(第7章「[ソフトウェアRAIDの設定](#)」を参照)。ただし、LVMはRAID 0および1では適切に動作します。これは、RAID 0は一般的な論理ボリューム管理と同様である(個々の論理ブロックが物理デバイス上のブロックにマップされる)ためです。RAID 1上でLVMを使用した場合は、ミラーの同期を追跡して同期プロセスを完全に管理することができます。それより高いRAIDレベルでは、接続されたディスクの状態を監視するほか、ディスクアレイで問題が発生した場合に管理者に通知することのできる、管理デーモンが必要になります。LVMにはこのようなデーモンが組み込まれていますが、デバイス障害などの例外的な状況では、このデーモンは正しく機能しません。



警告: IBM Z: LVMルートファイルシステム

LVMまたはソフトウェアRAIDアレイでルートファイルシステムを使用してシステムを設定する場合、`/boot`を別個の非LVMまたは非RAIDパーティションに配置する必要があります。そうしないと、システムは起動しません。このパーティションの推奨サイズは500MBで、推奨ファイルシステムはExt4です。

これらの機能とともにLVMを使用することは、頻繁に使用されるホームPCや小規模サーバではそれだけでも意義があります。データベース、音楽アーカイブ、またはユーザディレクトリのように増え続けるデータストックがある場合は、LVMが特に役に立ちます。LVMを使用すると、物理ハードディスクより大きなファイルシステムの作成が可能になります。ただし、LVMでの作業は従来のパーティションでの作業とは異なることに留意してください。

YaSTパーティショナの使用によって、新規および既存のLVMストレージオブジェクトを管理できます。LVMの設定に関する指示や詳細情報については、公式のLVM HOWTO (<https://tldp.org/HOWTO/LVM-HOWTO/>) を参照してください。

5.2 ボリュームグループの作成

LVMボリュームグループ(VG)は、Linux LVMパーティションをスペースの論理プールにします。グループ内の使用可能なスペースから論理ボリュームを作成できます。グループ内のLinux LVMパーティションは、同じディスクに存在することも、さまざまなディスクに存在することも可能です。パーティションまたはディスク全体を追加することにより、グループのサイズを拡張できます。

ディスク全体を使用する場合、そのディスクにパーティションを含めることはできません。パーティションを使用した場合、それらをマウントしないでください。YaSTは、パーティションをVGに追加する際に自動的にパーティションタイプを0x8E Linux LVMに変更します。

1. YaSTを起動してパーティショナを開きます。
2. 既存のパーティショニングセットアップを再設定する必要がある場合は、次の手順に従います。詳細については、『展開ガイド』、第11章「エキスパートパーティショナ」、11.1項「熟練者向けパーティション設定の使用」を参照してください。未使用的ディスクまたはパーティションを使用したいだけの場合は、この手順をスキップしてください。



警告: パーティションされていないディスクの物理ボリューム

パーティションされていないディスクがオペレーティングシステムのインストール先(ブート元)ではない場合、そのディスクを物理ボリューム(PV)として使用することができます。

パーティションされていないディスクはシステムレベルで「未使用」として表示されるため、上書きされてしまったり、間違ってアクセスされたりする可能性があります。

- a. 既にパーティションが含まれているハードディスク全体を使用するには、そのディスク上にあるパーティションをすべて削除します。
 - b. 現在マウントされているパーティションを使用するには、そのパーティションをアンマウントします。
3. 左のパネルで、ボリューム管理を選択します。
既存のボリュームグループのリストが右のパネルに表示されます。
4. [ボリューム管理] ページの左下で、ボリュームグループの追加をクリックします。



5. ボリュームグループは次のように定義します。

a. ボリュームグループ名を指定します。

インストール時にボリュームグループを作成している場合は、SUSE Linux Enterprise Serverのシステムファイルを含むボリュームグループに対して system という名前が示唆されます。

b. PEサイズを指定します。

PEサイズは、ボリュームグループの物理ブロックのサイズを定義します。ボリュームグループにある全ディスクスペースはこの物理ブロックサイズ内で使用されます。値の範囲は、2の累乗で1KBから16GBまでです。通常、この値は4MBに設定されます。

LVM1では、LVごとに65534エクステントまでしかサポートしないので、4MBの物理エクステントで最大LVサイズとして256GBが可能でした。SUSE Linux Enterprise Serverで使用されるLVM2では、物理エクステントの数に制限はありません。エクステントが多くても、論理ボリュームに対するI/Oパフォーマンスには影響しませんが、LVMツールの動作が遅くなります。



重要: 物理エクステントサイズ

1つのボリュームグループに異なるサイズの物理エクステントを混在させないでください。初期設定後はエクステントを変更しないでください。

- c. 利用可能な物理ボリュームリストで、このボリュームグループに含めたいLinux LVMパーティションを選択し、追加をクリックして、それらのパーティションを選択した物理ボリュームリストに移動します。
 - d. 完了をクリックします。
ボリュームグループリストに新しいグループが表示されます。
6. [ボリューム管理] ページで、次へをクリックし、新しいグループが一覧されることを確認してから、完了をクリックします。
7. ボリュームグループを構成している物理デバイスを確認するため、稼働中のシステムでYaSTパーティショナを開き、ボリューム管理 > 編集 > Physical Devices (物理デバイス) の順にクリックします。中止するをクリックしてこの画面を閉じます。

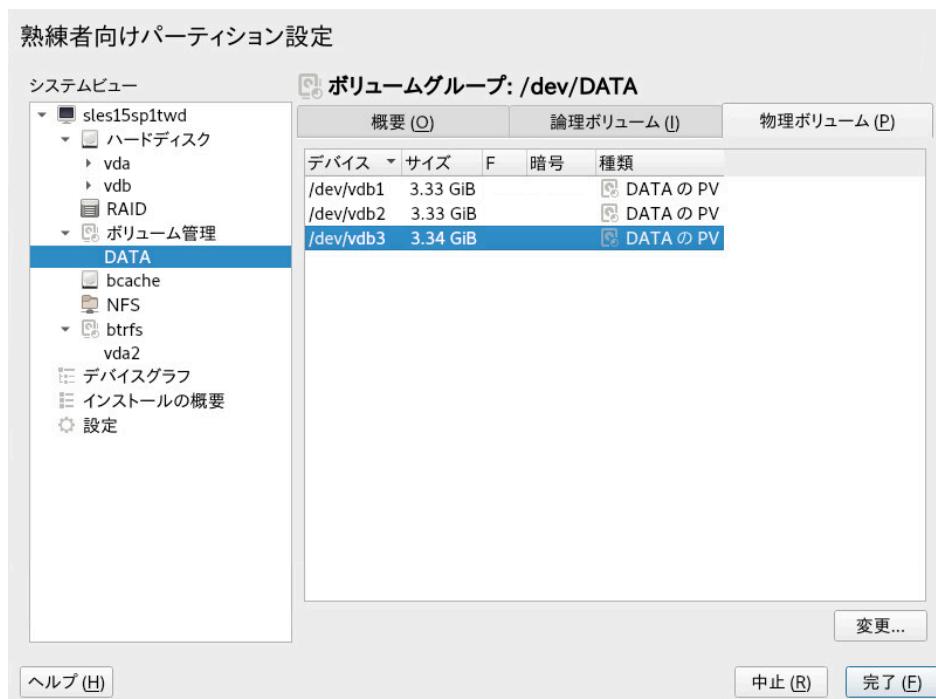


図 5.2: DATAという名前のボリュームグループ内の物理ボリューム

5.3 論理ボリュームの作成

論理ボリュームは、ハードディスクと同様に領域のプールを提供します。この領域を使用可能にするには、論理ボリュームを定義する必要があります。論理ボリュームは通常のパーティションに似ており、フォーマットやマウントが可能です。

YaSTパーティショナを使用して、既存のボリュームグループから論理ボリュームを作成します。各ボリュームグループに少なくとも1つの論理ボリュームを割り当ててください。ボリュームグループ内の空き領域を使い果たすまで、必要に応じて新しい論理ボリュームを作成できます。LVM論理ボリュームをオプションでシンプロビジョニングすることによって、使用可能な空き領域を超えるサイズで論理ボリュームを作成することもできます(詳しくは5.3.1項「シンプロビジョニング論理ボリューム」を参照)。

- **通常のボリューム:** (デフォルト)ボリュームの領域は直ちに割り当てられます。
- **シンプール:** この論理ボリュームは、シンボリューム用に予約された領域のプールです。シンボリュームでは、必要な領域をそのプールからオンデマンドで割り当てることができます。
- **シンボリューム:** ボリュームは疎ボリュームとして作成されます。このボリュームでは、必要な領域はシンプールからオンデマンドで割り当てられます。
- **ミラーリングされたボリューム:** このボリュームは、定義した数のミラーで作成されます。

手順 5.1: 論理ボリュームの設定

1. YaSTを起動してパーティショナを開きます。
2. 左のパネルで、ボリューム管理を選択します。既存のボリュームグループのリストが右のパネルに表示されます。
3. ボリュームを作成するボリュームグループを選択して、論理ボリューム、論理ボリュームの追加の順に選択します。
4. 名前にボリューム名を入力し、通常ボリュームを選択します(シンプロビジョニングボリュームの設定については、5.3.1項「シンプロビジョニング論理ボリューム」を参照してください)。次へで続行します。

論理ボリュームを DATA に追加



5. ボリュームのサイズと、複数ストライプを使用するかどうかを指定します。

ストライプボリュームを使用すると、データは複数の物理ボリュームに分散されます。これらの物理ボリュームが別のハードディスクに存在する場合、この性質により、読み込みおよび書き込みのパフォーマンスが向上します(RAID 0など)。利用可能な最大ストライプ数は、物理ボリュームの数と同じです。デフォルト(1)は、複数のストライプを使用しない設定です。

論理ボリュームを DATA に追加

サイズ

最大サイズ (9.99 GiB)
 カスタムサイズ
サイズ
9.99 GiB

ストライプ

数 サイズ
1 4 KiB

ヘルプ (H)

キャンセル (C) 戻る (B) 次へ (N)

6. 役割でボリュームの役割を選択します。ここで選択した内容は、次のダイアログのデフォルト値にのみ影響します。値は次の手順で変更可能です。わからない場合は、RAW ボリューム(未フォーマット)を選択します。

論理ボリュームを DATA に追加

役割

オペレーティングシステム
 データおよび CSV アプリケーション
 スwap
 EFI 起動パーティション
 何もしないボリューム (フォーマットしない)

ヘルプ (H)

キャンセル (C) 戻る (B) 次へ (N)

7. フォーマットオプションで、パーティションをフォーマットするを選択し、ファイルシステムを選択します。オプションメニューの内容は、ファイルシステムによって異なります。通常は、デフォルト値を変更する必要はありません。
マウントのオプションの下で、パーティションをマウントするを選択してから、マウントポイントを選択します。Fstabオプションをクリックして、このボリュームの特別なマウントオプションを追加します。
8. 完了をクリックします。
9. 次へをクリックし、変更が一覧されることを確認してから、完了をクリックします。

5.3.1 シンプロビジョニング論理ボリューム

LVM論理ボリュームはシンプロビジョニング可能です(オプション)。シンプロビジョニングを使用すると、利用可能な空き領域を超えるサイズの論理ボリュームを作成できます。任意の数のシンボリューム用に予約した未使用領域が含まれるシンプールを作成します。シンボリュームは疎ボリュームとして作成され、必要に応じてシンプールから領域が割り当てられます。ストレージ領域をコスト効果の高い方法で割り当たなければならなくなつた場合、シンプールを動的に拡張できます。シンプロビジョニングボリュームは、Snapperで管理可能なスナップショットもサポートします。詳細については、『管理ガイド』、第10章「Snapperを使用したシステムの回復とスナップショット管理」を参照してください。

シンプロビジョニング論理ボリュームを設定するには、[手順5.1 「論理ボリュームの設定」](#) の説明に従って作業を進めます。ボリュームタイプを選択する手順になつたら、通常ボリュームを選択せずに、シンボリュームまたはシンプールを選択します。

シンプール

この論理ボリュームは、シンボリューム用に予約された領域のプールです。シンボリュームでは、必要な領域をそのプールからオンデマンドで割り当てる事ができます。

シンボリューム

ボリュームは疎ボリュームとして作成されます。このボリュームでは、必要な領域はシンプールからオンデマンドで割り当たされます。

! 重要: クラスタにおけるシンプロビジョニングボリューム

クラスタでシンプロビジョニングボリュームを使用するには、クラスタを使用するシンプールとシンボリュームを1つのクラスタリソースで管理する必要があります。これにより、シンボリュームとシンプールを常に同じノードに排他的にマウントできます。

5.3.2 ミラーリングされたボリュームの作成

複数のミラーを使用して1つの論理ボリュームを作成できます。LVMは、下層の物理ボリュームに書き込まれたデータが別の物理ボリュームに確実にミラーリングされるようにします。そのため、1つの物理ボリュームがクラッシュしても、論理ボリューム上のデータにアクセスできます。LVMは、同期プロセスを管理するためのログファイルも保持します。このログには、現在ミラーとの同期を実行中のボリューム領域についての情報が含まれます。デフォルトでは、ログはディスク(可能であればミラーとは別のディスク)に保存されます。ただし、揮発性メモリなどの別の場所をログに指定できます。

現在のところ、使用可能なミラー実装のタイプには、「通常」(非RAID)のmirror論理ボリュームと、raid1論理ボリュームがあります。

ミラーリングされた論理ボリュームを作成したら、ミラーリングされた論理ボリュームで、アクティブ化、拡張、削除などの標準の操作を実行できます。

5.3.2.1 ミラーリングされた非RAID論理ボリュームの設定

ミラーリングされたボリュームを作成するには、lvcreateコマンドを使用します。次の例では、ボリュームグループ「vg1」を使用する、「lv1」という名前の2つのミラーを使用して、500GBの論理ボリュームを作成しています。

```
> sudo lvcreate -L 500G -m 2 -n lv1 vg1
```

このような論理ボリュームは、ファイルシステムのコピーを3つ提供するリニアボリューム(ストライピングなし)です。mオプションは、ミラーの数を指定します。Lオプションは、論理ボリュームのサイズを指定します。

論理ボリュームは、デフォルトサイズである512KBの領域に分割されます。異なるサイズの領域が必要な場合は、-Rオプションを使用します。このオプションの後に、目的の領域サイズをメガバイト単位で指定してください。または、mirror_region_sizeファイルのlvm.confオプションを編集して、好みの領域サイズを設定することもできます。

5.3.2.2 raid1論理ボリュームの設定

LVMはRAIDをサポートしているため、RAID1を使用してミラーリングを実装できます。このような実装には、非RAIDミラーと比較して次のような利点があります。

- LVMは、各ミラーイメージに対して完全に冗長なビットマップ領域を維持しており、これによって障害対応能力が向上する。
- ミラーイメージを一時的にアレイから分離し、マージして元に戻すことができる。

- 一時的な障害にアレイで対応できる。
- LVMのRAID 1実装はスナップショットをサポートする。

一方、このタイプのミラーリング実装では、クラスタ化されたボリュームグループ内に論理ボリュームを作成することはできません。

RAIDを使用してミラー ボリュームを作成するには、次のコマンドを発行します。

```
> sudo lvcreate --type raid1 -m 1 -L 1G -n lv1 vgl
```

各オプション/パラメータには次のような意味があります。

- --type - raid1を指定する必要があります。指定しないと、暗黙のセグメントタイプmirrorが使用され、非RAIDミラーが作成されます。
- -m - ミラーの数を指定します。
- -L - 論理ボリュームのサイズを指定します。
- -n - このオプションを使用して、論理ボリュームの名前を指定します。
- vgl - 論理ボリュームで使用されるボリュームグループの名前です。

LVMは、アレイ内の各データボリュームに対して、1つのエクステントサイズの論理ボリュームを作成します。ミラーリングされたボリュームが2つある場合、LVMは、メタデータを保存する別のボリュームを2つ作成します。

RAID論理ボリュームを作成したら、一般的な論理ボリュームと同じ方法でそのボリュームを使用できます。アクティブ化、拡張などを行うことができます。

5.4 非ルートLVMボリュームグループの自動アクティブ化

非ルートLVMボリュームグループのアクティブ化の動作は、/etc/lvm/lvm.confファイルとauto_activation_volume_listパラメータで制御されます。デフォルトでは、このパラメータは空で、すべてのボリュームがアクティブ化されます。一部のボリュームグループのみをアクティブ化するには、その名前を引用符で囲んで追加し、カンマで区切れます。次に例を示します。

```
auto_activation_volume_list = [ "vg1", "vg2/lvol1", "@tag1", "@*" ]
```

リストをauto_activation_volume_listパラメータで定義した場合、次のように処理されます。

1. 各論理ボリュームは、最初にこのリストに照らして確認されます。
2. 一致しない場合、論理ボリュームはアクティブ化されません。

デフォルトでは、非ルートLVMボリュームグループは、システムの再起動時にdracutによって自動的にアクティブ化されます。このパラメータにより、システムの再起動時にすべてのボリュームグループをアクティブにすることも、または指定した非ルートLVMボリュームグループのみをアクティブにすることもできます。

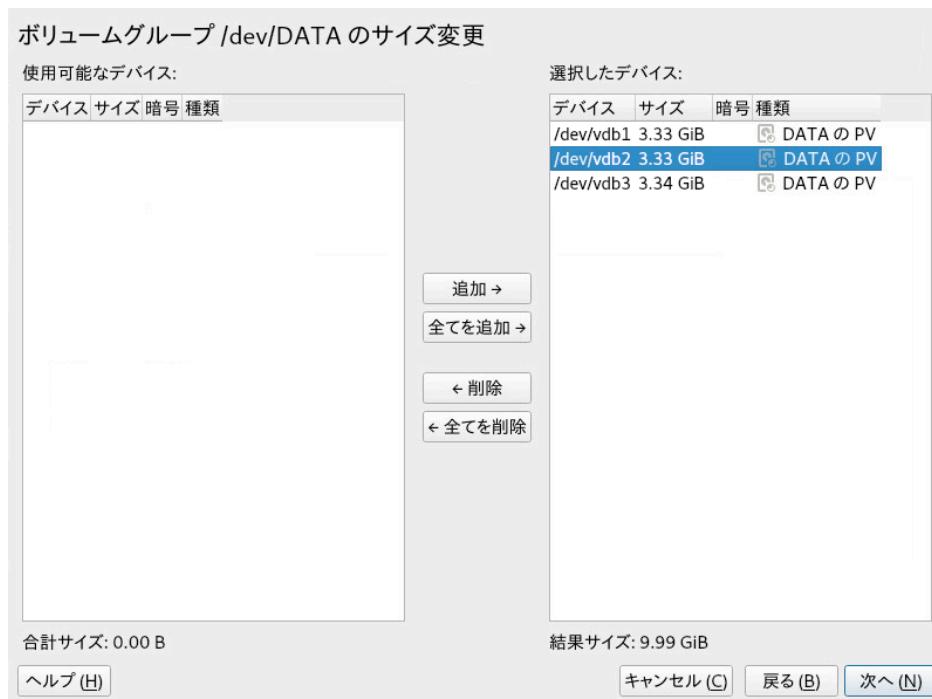
5.5 既存のボリュームグループのサイズ変更

ボリュームグループによって提供される領域は、物理ボリュームを追加することによっていつでも拡張できます。これは、システムの稼働中であっても、サービスを中断することなく実行できます。これにより、グループに論理ボリュームを追加したり、既存のボリュームのサイズを拡張したりできます。[5.6項 「論理ボリュームのサイズ変更」](#)を参照してください。

また、物理ボリュームを削除してボリュームグループのサイズを縮小することもできます。YaSTで削除できる物理ボリュームは、現在未使用の物理ボリュームだけです。現在使用中の物理ボリュームを確認するには、次のコマンドを実行します。PE Ranges列に表示されているパーティション(物理ボリューム)が使用中のものです。

```
> sudo pvs -o vg_name,lv_name,pv_name,seg_pe_ranges
root's password:
  VG   LV   PV          PE Ranges
            /dev/sda1
  DATA  DEVEL /dev/sda5  /dev/sda5:0-3839
  DATA        /dev/sda5
  DATA  LOCAL /dev/sda6  /dev/sda6:0-2559
  DATA        /dev/sda7
  DATA        /dev/sdb1
  DATA        /dev/sdc1
```

1. YaSTを起動してパーティショナを開きます。
2. 左のパネルで、ボリューム管理を選択します。既存のボリュームグループのリストが右のパネルに表示されます。
3. 変更するボリュームグループを選択し、物理ボリュームタブを有効にして、変更をクリックします。



4. 次のいずれかの操作を行います。

- **追加:** 1つまたは複数の物理ボリューム(LVMパーティション)を利用可能な物理ボリュームリストから選択した物理ボリュームリストに移動することにより、ボリュームグループのサイズを拡張します。
- **削除:** 1つまたは複数の物理ボリューム(LVMパーティション)を選択した物理ボリュームリストから使用可能な物理ボリュームリストに移動することにより、ボリュームグループのサイズを縮小します。

5. 完了をクリックします。

6. 次へをクリックし、変更が一覧されることを確認してから、完了をクリックします。

5.6 論理ボリュームのサイズ変更

ボリュームグループ内に利用可能な未使用の空き領域がある場合、論理ボリュームを拡張して使用可能な領域を増やすことができます。また、ボリュームのサイズを縮小してボリュームグループの領域を解放し、他の論理ボリュームで使用できるようにすることもできます。



注記: 「オンライン」でのサイズ変更

ボリュームのサイズを縮小すると、そのファイルシステムのサイズもYaSTによって自動的に縮小されます。現在マウントされているボリュームのサイズを「オンライン」で(つまりマウント中に)変更できるかどうかは、ファイルシステムによって異なります。オンライン拡張をサポートするファイルシステムは、Btrfs、XFS、Ext3、およびExt4です。

オンライン縮小をサポートするファイルシステムは、Btrfsのみです。Ext2/3/4ファイルシステムを縮小するには、アンマウントする必要があります。XFSはファイルシステムの縮小をサポートしないため、XFSでフォーマットされたボリュームは縮小できません。

1. YaSTを起動して、システムの下にあるパーティショナを開きます。
2. 左パネルで、LVMボリュームグループを選択します。
3. 右パネルで変更する論理ボリュームを選択します。
4. デバイス、サイズ変更をクリックします。



5. 次のオプションの1つを使用して目的のサイズを設定します。

- **最大サイズ.** 論理ボリュームのサイズを、ボリュームグループの残り領域をすべて使用するよう拡張します。
- **最小サイズ.** 論理ボリュームのサイズを、データおよびファイルシステムメタデータによって使用されているサイズまで縮小します。
- **カスタムサイズ.** ボリュームの新しいサイズを指定します。上に表示されている最小値から最大値までの範囲内の値を指定する必要があります。キロバイトにはK、メガバイトにはM、ギガバイトにはG、テラバイトにはTをそれぞれ使用します(たとえば20G)。

6. 次へをクリックし、変更が一覧されることを確認してから、完了をクリックします。

5.7 ボリュームグループまたは論理ボリュームの削除



警告: データ損失

ボリュームグループを削除すると、グループの各メンバーパーティションに含まれているデータがすべて破棄されます。論理ボリュームを削除すると、そのボリュームに保存されているデータがすべて破棄されます。

1. YaSTを起動してパーティショナを開きます。
2. 左のパネルで、ボリューム管理を選択します。既存のボリュームグループのリストが右のパネルに表示されます。
3. 削除するボリュームグループまたは論理ボリュームを選択して、Delete (削除)をクリックします。
4. 選択した内容に応じて警告ダイアログが表示されます。はいを選択して確認します。
5. 次へをクリックして、削除されたボリュームグループが一覧表示されていることを確認します。削除は赤色フォントで示されます。完了をクリックします。

5.8 起動時のLVMの無効化

LVMストレージでエラーが発生した場合、LVMボリュームをスキャンすると、緊急/レスキュー・シェルが起動しなくなる場合があります。その結果、問題を詳細に診断できなくなります。LVMストレージで障害が発生した場合にこのスキャンを無効にするために、カーネルコマンドラインの`nolvmd`オプションを渡すことができます。

5.9 LVMコマンドの使用

LVMコマンドの使用の詳細については、次の表で説明されている各コマンドの`man`ページを参照してください。すべてのコマンドは`root`特権で実行する必要があります。`sudo COMMAND`を使用するか(推奨)、直接`root`として実行します。

LVMコマンド

`pvccreate DEVICE`

LVMで物理ボリュームとして使用できるようにデバイス(`/dev/sdb1`など)を初期化します。指定したデバイス上にファイルシステムが存在する場合、警告が表示されます。`blkid`がインストールされている場合にのみ(デフォルトでインストールされています)、`pvccreate`により既存のファイルシステムの有無が確認されることを覚えておいてください。`blkid`が使用可能でない場合、`pvccreate`によって何も警告が生成されず、警告なしにファイルシステムが失われる場合があります。

`pvdisplay DEVICE`

LVM物理ボリュームに関する情報(現在、論理ボリュームで使用中かどうかなど)を表示します。

`vgcreate -c y VG_NAME DEV1 [DEV2...]`

指定した1つ以上のデバイスでクラスタ化ボリュームグループを作成します。

`vgcreate --activationmode ACTIVATION_MODE VG_NAME`

ボリュームグループのアクティブ化のモードを設定します。次のいずれかの値を指定できます。

- complete - 欠落している物理ボリュームの影響を受けない論理ボリュームのみをアクティブ化できます。特定の論理ボリュームでそのような障害が許容される場合も、同様の処理が実行されます。
- degraded - デフォルトのアクティブ化モードです。論理ボリュームをアクティブ化するための十分なレベルの冗長性がある場合、一部の物理ボリュームが欠落しても、その論理ボリュームをアクティブ化できます。
- partial - LVMは、一部の物理ボリュームが欠落していても、ボリュームグループのアクティブ化を試みます。非冗長論理ボリュームから重要な物理ボリュームが欠落している場合、通常、その論理ボリュームはアクティブ化できず、エラーターゲットとして扱われます。

`vgchange -a [ey|n] VG_NAME`

ボリュームグループおよびその論理ボリュームを入出力用にアクティブ化(-a ey)または非アクティブ化(-a n)します。

クラスタ内のボリュームをアクティブ化する場合は、必ず ey オプションを使用してください。ロードスクリプトではこのオプションがデフォルトで使用されます。

`vgremove VG_NAME`

ボリュームグループを削除します。このコマンドを使用する前に、論理ボリュームを削除してボリュームグループを非アクティブにしてください。

`vgdisplay VG_NAME`

指定したボリュームグループに関する情報を表示します。

ボリュームグループの合計物理エクステントを確認するには、次のように入力します。

```
> vgdisplay VG_NAME | grep "Total PE"
```

`lvcreate -L SIZE -n LV_NAME VG_NAME`

指定したサイズの論理ボリュームを作成します。

`lvcreate -L SIZE --thinpool POOL_NAME VG_NAME`

ボリュームグループ VG_NAME から、指定したサイズのシンプール myPool を作成します。

次の例では、ボリュームグループ LOCAL から 5GB のサイズのシンプールを作成します。

```
> sudo lvcreate -L 5G --thinpool myPool LOCAL
```

`lvcreate -T VG_NAME/POOL_NAME -V SIZE -n LV_NAME`

プール POOL_NAME 内にシンボリカルボリュームを作成します。次の例では、ボリュームグループ LOCAL 上のプール myPool から 1GB のシンボリカルボリューム myThin1 を作成します。

```
> sudo lvcreate -T LOCAL/myPool -V 1G -n myThin1
```

lvcreate -T VG_NAME/POOL_NAME -V SIZE -L SIZE -n LV_NAME

シンプールの作成とシン論理ボリュームの作成を1つのコマンドに結合することもできます。

```
> sudo lvcreate -T LOCAL/myPool -V 1G -L 5G -n myThin1
```

lvcreate --activationmode ACTIVATION_MODE LV_NAME

論理ボリュームのアクティブ化のモードを設定します。次のいずれかの値を指定できます。

- complete - 論理ボリュームは、そのすべての物理ボリュームがアクティブな場合にのみアクティブ化できます。
- degraded - デフォルトのアクティブ化モードです。論理ボリュームをアクティブ化するための十分なレベルの冗長性がある場合、一部の物理ボリュームが欠落しても、その論理ボリュームをアクティブ化できます。
- partial - LVMは、一部の物理ボリュームが欠落していても、ボリュームのアクティブ化を試みます。この場合、論理ボリュームの一部が使用できなくなり、データが消失することがあります。このオプションは通常は使用しませんが、データを復元する場合に役立つことがあります。

activation_mode設定オプションの上記いずれかの値を指定することによって、/etc/lvm/lvm.confでアクティブ化モードを指定することもできます。

lvcreate -s [-L SIZE] -n SNAP_VOLUME SOURCE_VOLUME_PATH VG_NAME

指定した論理ボリュームに対してスナップショットボリュームを作成します。サイズオプション(-Lまたは--size)を指定しなかった場合、スナップショットはシンスナップショットとして作成されます。

lvremove /dev/VG_NAME/LV_NAME

論理ボリュームを削除します。

このコマンドを使用する前に、論理ボリュームをumountコマンドでアンマウントして閉じてください。

lvremove SNAP_VOLUME_PATH

スナップショットボリュームを削除します。

lvconvert --merge SNAP_VOLUME_PATH

論理ボリュームをスナップショットのバージョンに戻します。

vgextend VG_NAME DEVICE

指定したデバイス(物理ボリューム)を既存のボリュームグループに追加します。

vgreduce VG_NAME DEVICE

指定した物理ボリュームを既存のボリュームグループから削除します。

物理ボリュームが論理ボリュームによって使用中でないことを確認してください。使用中の場合は、**pvmove**コマンドを使用してデータを別の物理ボリュームに移動する必要があります。

lvextend -L SIZE /dev/VG_NAME/LV_NAME

指定した論理ボリュームのサイズを拡張します。その後、新たに使用可能になった領域を使用するため、ファイルシステムを拡張する必要もあります。詳細については[第2章「ファイルシステムのサイズ変更」](#)を参照してください。

lvreduce -L SIZE /dev/VG_NAME/LV_NAME

指定した論理ボリュームのサイズを縮小します。

ボリュームを縮小する前に、まずファイルシステムのサイズを縮小してください。そうしないと、データを失うリスクがあります。詳細については[第2章「ファイルシステムのサイズ変更」](#)を参照してください。

lvrename /dev/VG_NAME/LV_NAME /dev/VG_NAME/NEW_LV_NAME

既存のLVM論理ボリュームの名前を変更します。ボリュームグループの名前は変更されません。



ヒント: ボリューム作成時のudevのバイパス

udevルールではなくLVMを使用してLVデバイスノードとシンボリックリンクを管理する場合は、次のいずれかの方法でudevからの通知を無効にすることによって可能になります。

- `/etc/lvm/lvm.conf`で、`activation/udev_rules = 0`と`activation/udev_sync = 0`を設定する。
`lvcreate`コマンドで`--nodevsync`を指定しても、`activation/udev_sync = 0`と同じ結果になります。この場合も、`activation/udev_rules = 0`の設定が必要です。
- 環境変数`DM_DISABLE_UDEV`を設定する。

```
export DM_DISABLE_UDEV=1
```

この方法でも、udevからの通知が無効になります。さらに、/etc/lvm/lvm.confのudev関連の設定はすべて無視されます。

5.9.1 コマンドによる論理ボリュームのサイズ変更

lvresize、lvextend、およびlvreduceコマンドが論理ボリュームのサイズ変更に使用されます。構文とオプションについては、これらの各コマンドのマニュアルページを参照してください。LVを拡大するには、VG上に十分な未使用スペースがなければなりません。

論理ボリュームを拡大または縮小する場合、YaSTパーティショナを使用することをお勧めします。YaSTを使用すると、そのボリュームのファイルシステムのサイズも自動的に調整されます。

LVは使用中に手動で拡大または縮小できますが、LV上のファイルシステムについてはこれが不可能な場合があります。LVを拡大、縮小しても、そのボリューム内のファイルシステムのサイズは自動的に変更されません。後でファイルシステムを拡大するには、別のコマンドを使用する必要があります。ファイルシステムのサイズ変更の詳細については、[第2章 「ファイルシステムのサイズ変更」](#)を参照してください。

手動でLVのサイズを変更する場合は、次に示すように正しい順序に従ってください。

- LVを拡大する場合は、ファイルシステムを拡大する前にLVを拡大する必要があります。
- LVを縮小する場合は、LVを縮小する前にファイルシステムを縮小する必要があります。

論理ボリュームのサイズを拡張するには:

1. 端末を開きます。
2. 論理ボリュームにExt2またはExt4ファイルシステム(オンライン拡張がサポートされていません)が含まれる場合、マウント解除します。仮想マシン(Xen VMなど)用に提供されているファイルシステムが含まれている場合は、最初にVMをシャットダウンします。
3. 端末のプロンプトに対して、次のコマンドを入力し、論理ボリュームのサイズを拡大します。

```
> sudo lvextend -L +SIZE /dev/VG_NAME/LV_NAME
```

SIZEの場合は、10GBのように、論理ボリュームに追加したい容量を指定してください。/dev/VG_NAME/LV_NAMEを、/dev/LOCAL/DATAなどの論理ボリュームへのLinuxパスに入れ替えます。例:

```
> sudo lvextend -L +10GB /dev/vg1/v1
```

4. ファイルシステムのサイズを調整します。詳細については[第2章「ファイルシステムのサイズ変更」](#)を参照してください。
5. ファイルシステムをマウント解除した場合は、再びマウントします。

たとえば、LVをLV上の(マウント済みでアクティブな)Btrfsで10GB拡張するには:

```
> sudo lvextend -L +10G /dev/LOCAL/DATA  
> sudo btrfs filesystem resize +10G /dev/LOCAL/DATA
```

論理ボリュームのサイズを縮小するには:

1. 端末を開きます。
2. 論理ボリュームにBtrfsファイルが含まれていない場合は、論理ボリュームをマウント解除します。仮想マシン(Xen VMなど)用に提供されているファイルシステムが含まれている場合は、最初にVMをシャットダウンします。XFSファイルシステムを使用しているボリュームのサイズは縮小できません。
3. ファイルシステムのサイズを調整します。詳細については[第2章「ファイルシステムのサイズ変更」](#)を参照してください。
4. 端末のプロンプトに対して、次のコマンドを入力し、論理ボリュームのサイズをファイルシステムのサイズまで縮小します。

```
> sudo lvreduce /dev/VG_NAME/LV_NAME
```

5. ファイルシステムをアンマウントしてあった場合は、再びマウントします。

たとえば、LVをLV上のBtrfsで5GB縮小するには:

```
> sudo btrfs filesystem resize -size 5G /dev/LOCAL/DATA  
sudo lvreduce /dev/LOCAL/DATA
```



ヒント: 1つのコマンドでのボリュームとファイルシステムのサイズ変更

SUSE Linux Enterprise Server 12 SP1から、`lvextend`、`lvresize`、および`lvreduce`で`--resizesfs`オプションがサポートされるようになりました。このオプションは、ボリュームのサイズを変更するだけでなく、ファイルシステムのサイズも変更します。したがって、上に示す`lvextend`および`lvreduce`の例は、次のように実行することもできます。

```
> sudo lvextend --resizesfs -L +10G /dev/LOCAL/DATA  
> sudo lvreduce --resizesfs -L -5G /dev/LOCAL/DATA
```

`--resizesfs`は、ext2/3/4、Btrfs、およびXFSの各ファイルシステムでサポートされます。このオプションを使用したBtrfsのサイズ変更は、まだ上流では許可されていないため、SUSE Linux Enterprise Serverでのみ可能です。

5.9.2 LVMキャッシュボリュームの使用

LVMでは、大容量の低速なブロックデバイスに対して、高速なブロックデバイス(SSDデバイスなど)をライトバックキャッシュまたはライトスルーキャッシュとして使用できます。キャッシュ論理ボリュームタイプは、小容量の高速なLVを使用して、大容量の低速なLVのパフォーマンスを向上させます。

LVMキャッシングを設定するには、キャッシングデバイス上に2つの論理ボリュームを作成する必要があります。大容量の論理ボリュームはキャッシング自体に使用され、小容量のボリュームはキャッシングメタデータの保存に使用されます。これら2つのボリュームは、元のボリュームと同じボリュームグループに属している必要があります。これらのボリュームを作成したら、キャッシングプールに変換して元のボリュームに接続する必要があります。

手順 5.2: キャッシュ論理ボリュームの設定

1. 元のボリュームがまだ存在しない場合は(低速なデバイス上に)作成します。
2. 物理ボリュームを(高速なデバイスから)元のボリュームが属するボリュームグループに追加して、物理ボリューム上にキャッシングデータボリュームを作成します。
3. キャッシュメタデータボリュームを作成します。サイズは、キャッシングデータボリュームの1/1000にする必要があります。最小サイズは8MBです。
4. キャッシュデータボリュームとメタデータボリュームをキャッシングプールボリュームに結合します。

```
> sudo lvconvert --type cache-pool --poolmetadata VOLUME_GROUP/
METADATA_VOLUME VOLUME_GROUP/CACHING_VOLUME
```

5. キャッシュプールを元のボリュームに接続します。

```
> sudo lvconvert --type cache --cachepool VOLUME_GROUP/CACHING_VOLUME VOLUME_GROUP/
ORIGINAL_VOLUME
```

LVMキャッシングの詳細については、lvmcache(7)のマニュアルページを参照してください。

5.10 LVM2ストレージオブジェクトへのタグ付け

タグは、ストレージオブジェクトのメタデータに割り当たる順序付けのないキーワードまたは用語です。タグを使用すると、順序付けのないタグのリストをLVMストレージオブジェクトのメタデータに添付することによって、それらのオブジェクトのコレクションを有用なように分類できます。

5.10.1 LVM2タグの使用

LVM2ストレージオブジェクトにタグを付けたら、それらのタグをコマンドで使用して、次のタスクを達成できます。

- 特定のタグの有無に応じて、処理するLVMオブジェクトを選択します。
- 設定ファイル内でタグを使用することにより、サーバ上でアクティブにするボリュームグループと論理ボリュームを制御します。
- コマンド内でタグを指定することにより、グローバル設定ファイルの設定を上書きします。

コマンドラインでLVMオブジェクトを参照する代わりに、タグを使用して、次の項目を受け入れることができます。

- オブジェクトのリスト
- 单一のオブジェクト(タグが單一オブジェクトに展開する限り)

オブジェクト名をタグで置き換えることは、一部ではサポートされていません。引数の展開後、リスト内の重複引数は、重複引数を削除し、各引数の最初のインスタンスを保留することによって解決されます。

引数のタイプが曖昧になる可能性がある場合は、タグの前にアットマーク(@)文字を付けてください(たとえば、@mytag)。それ以外の接頭辞「@」の使用はオプションです。

5.10.2 LVM2タグの作成要件

LVMでタグを使用する場合は、以下の要件を考慮してください。

サポートされている文字

LVMタグのワードには、ASCII 大文字A～Z、小文字a～z、数字0～9、下線(_)、プラス(+)、ハイフン(-)、およびピリオド(.)を含めることができます。ワードをハイフンで始めることはできません。最大128文字まで入力できます。

サポートされているストレージオブジェクト

タグ付けできるのは、LVM2の物理ボリューム、ボリュームグループ、論理ボリューム、および論理ボリュームセグメントです。PVタグは、そのボリュームグループのメタデータに保存されます。ボリュームグループを削除すると、孤立した物理ボリューム内のタグも削除されます。スナップショットにはタグを付けられませんが、元のオブジェクトはタグ付けできます。

LVM1オブジェクトは、そのディスクフォーマットがタグをサポートしていないので、タグ付けできません。

5.10.3 コマンドラインでのタグ構文

--addtagTAG_INFO

LVM2ストレージオブジェクトにタグを追加(つまり、**タグ付け**)します。例:

```
> sudo vgchange --addtag @db1 vgl
```

--deltagTAG_INFO

LVM2ストレージオブジェクトからタグを削除(つまり、**タグ解除**)します。例:

```
> sudo vgchange --deltag @db1 vgl
```

--tagTAG_INFO

アクティブまたは非アクティブにするボリュームグループまたは論理ボリュームのリストを絞り込むために使用するタグを指定します。

次の例に示すコマンドを入力すると、指定のタグに一致するタグをもつボリュームがアクティブになります。

```
> sudo lvchange -ay --tag @db1 vg1/vol2
```

5.10.4 設定ファイル構文

以降の各項では、特定の事例における設定例を示します。

5.10.4.1 lvm.confファイルでのホスト名タグの有効化

次のコードを`/etc/lvm/lvm.conf`ファイルに追加することにより、`/etc/lvm/lvm_<HOSTNAME>.conf`ファイルでホストに個別に定義されているホストタグを有効にします。

```
tags {  
    # Enable hostname tags  
    hosttags = 1  
}
```

ホストの`/etc/lvm/lvm_<HOSTNAME>.conf`ファイルにアクティベーションコードを入力します。5.10.4.3項「[アクティベーションを定義する](#)」を参照してください。

5.10.4.2 lvm.confファイルでホスト名タグを定義する

```
tags {  
  
    tag1 {}  
        # Tag does not require a match to be set.  
  
    tag2 {  
        # If no exact match, tag is not set.  
        host_list = [ "hostname1", "hostname2" ]  
    }  
}
```

5.10.4.3 アクティベーションを定義する

`/etc/lvm/lvm.conf`ファイルを変更すると、タグに基づいてLVM論理ボリュームをアクティブにできます。

テキストエディタで、次のコードをファイルに追加します。

```
activation {
    volume_list = [ "vg1/lvol0", "@database" ]
}
```

@databaseをご使用のタグで置き換えます。ホストに設定されているすべてのタグにタグを一致させるには、"@*"を使用します。

アクティベーションコマンドは、ボリュームグループと論理ボリュームのメタデータで設定されているVGNAME、VGNAME/LVNAME、または@TAGと照合を行います。ボリュームグループまたは論理グループは、メタデータタグが一致する場合のみアクティブになります。一致しない場合、デフォルトではアクティブになりません。

volume_listが存在せず、ホストにタグが定義されていると、ホストタグがメタデータタグに一致する場合のみボリュームグループまたは論理グループがアクティブになります。

volume_listが定義されていても空であり、ホストにタグが定義されていないと、アクティブなりません。

volume_listが定義されていないと、LVのアクティブ化に制限は課されません(すべて許可されます)。

5.10.4.4 複数のホスト名設定ファイルでアクティベーションを定義する

lvm.confファイルでホストタグが有効になっている場合、ホストの設定ファイル(/etc/lvm/lvm_<HOST_TAG>.conf)でアクティベーションコードを使用できます。たとえば、サーバの/etc/lvm/ディレクトリに、2つの設定ファイルがあるとします。

lvm.conf
lvm_<HOST_TAG>.conf

スタートアップ時に、/etc/lvm/lvm.confファイルがロードされ、ファイル内のすべてのタグ設定が処理されます。ホストタグが定義されている場合、関連する/etc/lvm/lvm_<HOST_TAG>.confファイルがロードされます。特定の設定ファイルエントリを検索する際、最初にホストタグファイルが検索されます。続いてlvm.confファイルが検索され、最初に一致した箇所で停止します。lvm_<HOST_TAG>.confファイル内で、タグが設定された順序とは逆の順序を使用します。これによって、最後に設定されたタグのファイルが最初に検索されます。ホストタグファイルで新しいタグが設定されると、追加の設定ファイルがロードされます。

5.10.5 クラスタで簡単なアクティベーション制御にタグを使用する

簡単なホスト名のアクティベーション制御は、/etc/lvm/lvm.confファイルでhostname_tagsオプションを有効にすることで設定できます。これがグローバル設定になるように、同じファイルをクラスタ内のですべてのコンピュータで使用します。

1. テキストエディタで、次のコードを/etc/lvm/lvm.confファイルに追加します。

```
tags {  
    hostname_tags = 1  
}
```

2. ファイルをクラスタ内のすべてのホストに複製します。
3. クラスタ内の任意のコンピュータから、vg1/lvol2をアクティブにするコンピュータのリストにdb1を追加します。

```
> sudo lvchange --addtag @db1 vg1/lvol2
```

4. db1サーバで、次のコードを入力してvg1/lvol2をアクティブにします。

```
> sudo lvchange -ay vg1/vol2
```

5.10.6 タグを使用して、クラスタ内好みのホストでアクティブにする

本項の例では、次のようなアクティベーションを行う2つの方法を示します。

- ボリュームグループvg1をデータベースホストdb1とdb2のみでアクティブにします。
- ボリュームグループvg2をファイルサーバホストfs1のみでアクティブにします。
- ファイルサーバのバックアップホストfsb1では、最初は何もアクティブにせず、ファイルサーバのホストfs1に置き換わる準備をします。

5.10.6.1 オプション1:一元化された管理とホスト間で複製された静的設定

次のソリューションでは、单一の設定ファイルを複数のホスト間で複製します。

1. @databaseタグをボリュームグループvg1のメタデータに追加します。端末で、次のコマンドを入力します。

```
> sudo vgchange --addtag @database vg1
```

2. @fileserverタグをボリュームグループvg2のメタデータに追加します。端末で、次のコマンドを入力します。

```
> sudo vgchange --addtag @fileserver vg2
```

3. テキストエディタで、次のコードを使用して /etc/lvm/lvm.confファイルを変更することにより、@database、@fileserver、@fileserverbackupの各タグを定義します。

```
tags {  
    database {  
        host_list = [ "db1", "db2" ]  
    }  
    fileserver {  
        host_list = [ "fs1" ]  
    }  
    fileserverbackup {  
        host_list = [ "fsb1" ]  
    }  
}  
  
activation {  
    # Activate only if host has a tag that matches a metadata tag  
    volume_list = [ "@*" ]  
}
```

4. 変更した /etc/lvm/lvm.confファイルを4つのホスト(db1、db2、fs1、およびfsb1)に複製します。
5. ファイルサーバホストが故障した場合は、次のコマンドを任意のモードで端末から入力することにより、fsb1上でvg2を起動できます。

```
> sudo vgchange --addtag @fileserverbackup vg2  
> sudo vgchange -ay vg2
```

5.10.6.2 オプション2: ローカライズされた管理と設定

次のソリューションでは、各ホストがアクティブにするボリュームのクラスに関する情報をローカルに保持します。

1. @databaseタグをボリュームグループvg1のメタデータに追加します。端末で、次のコマンドを入力します。

```
> sudo vgchange --addtag @database vg1
```

2. @fileserverタグをボリュームグループvg2のメタデータに追加します。端末で、次のコマンドを入力します。

```
> sudo vgchange --addtag @fileserver vg2
```

3. /etc/lvm/lvm.confファイルでホストタグを有効にします。

- a. テキストエディタで、次のコードを使用して/etc/lvm/lvm.confファイルを変更することにより、ホストタグ設定ファイルを有効にします。

```
tags {  
    hosttags = 1  
}
```

- b. 変更した/etc/lvm/lvm.confファイルを4つのホスト(db1、db2、fs1、およびfsb1)に複製します。

4. ホストdb1で、データベースホストdb1のアクティベーション設定ファイルを作成します。テキストエディタで、/etc/lvm/lvm_db1.confファイルを作成し、次のコードを追加します。

```
activation {  
    volume_list = [ "@database" ]  
}
```

5. ホストdb2で、データベースホストdb2のアクティベーション設定ファイルを作成します。テキストエディタで、/etc/lvm/lvm_db2.confファイルを作成し、次のコードを追加します。

```
activation {  
    volume_list = [ "@database" ]  
}
```

6. ホストfs1で、ファイルサーバホストfs1のアクティベーション設定ファイルを作成します。テキストエディタで、/etc/lvm/lvm_fs1.confファイルを作成し、次のコードを追加します。

```
activation {  
    volume_list = [ "@fileserver" ]
```

```
}
```

7. ファイルサーバホストfs1が故障した場合は、スペアのファイルサーバホストfsb1をファイルサーバとして起動します。

- a. ホストfsb1で、ホストfsb1のアクティベーション設定ファイルを作成します。テキストエディタで、/etc/lvm/lvm_fsb1.confファイルを作成し、次のコードを追加します。

```
activation {
    volume_list = [ "@fileserver" ]
}
```

- b. 端末で、次のコマンドの1つを入力します。

```
> sudo vgchange -ay vg2
> sudo vgchange -ay @fileserver
```

6 LVMボリュームスナップショット

LVM (Logical Volume Manager)論理ボリュームスナップショットはコピーオンライト技術の1つで、既存のボリュームのデータブロックに対する変更を監視し、いずれかのブロックに書き込みが行われると、スナップショット時のブロックの値がスナップショットボリュームにコピーされます。こうすることで、スナップショットボリュームが削除されるまで、データのその時点のコピーが保存されます。

6.1 ボリュームスナップショットの理解

ファイルシステムのスナップショットには、それ自体のメタデータと、スナップショットの作成後に変更されたソース論理ボリュームのデータブロックが含まれています。スナップショットを介してデータにアクセスすると、ソース論理ボリュームのその時点のコピーが表示されます。バックアップ媒体からデータを復元したり、変更されたデータを上書きする必要はありません。

! 重要: スナップショットによるボリュームのマウント

スナップショットのライフタイム中は、スナップショットを先にマウントしないと、ソース論理ボリュームをマウントできません。

LVMボリュームスナップショットでは、ファイルシステムのその時点のビューからバックアップを作成できます。スナップショットは瞬時に作成され、削除するまで保存されます。ボリューム自体はユーザが引き続き利用できるようにしながら、スナップショットからファイルシステムのバックアップを作成できます。当初のスナップショットには、スナップショットに関するメタデータが含まれていますが、ソース論理ボリュームの実際のデータは含まれていません。スナップショットはコピーオンライト技術を使用して、オリジナルデータブロックのデータ変更を検出します。スナップショットをとった際に保存されていた値をスナップショットボリューム内のブロックにコピーし、ソースブロックに新しいデータを保存することができます。ソース論理ボリュームで元の値から変更されるブロックが増えると、スナップショットのサイズが増えます。

スナップショットのサイズを決定する際には、ソース論理ボリュームに対して予想されるデータ変更量、およびスナップショットの保存期間を考慮する必要があります。スナップショットボリュームに割り当てるスペースの量は、ソース論理ボリュームのサイズ、スナップショットの保持予定期間、およびスナップショットのライフタイム中に変更が予期されるデータブロックの数によって異なります。スナップショットボリュームは、作成後のサイズ変更はできません。目安として、元の論理ボリュームの約10%のサイズで、スナップショットボ

リュームを作成してください。スナップショットの削除前に、ソース論理ボリューム内のすべてのブロックが1回以上変更されると予期される場合は、スナップボリュームのサイズを、少なくともソース論理ボリュームサイズにそのボリュームに関するメタデータ用スペースを加えたサイズにする必要があります。データ変更が頻繁でないか、またはライフタイムが十分短いと予期される場合、必要なスペースは少なくなります。

LVM2では、スナップショットはデフォルトで読み書き可能です。データをスナップショットに直接書き込む際は、そのブロックは例外テーブルで使用中とマークされ、ソース論理ボリュームからのコピーは行われません。スナップショットボリュームをマウントし、そのスナップショットボリュームにデータを直接書き込むことによって、アプリケーションの変更をテストできます。スナップショットをマウント解除してスナップショットを削除し、ソース論理ボリュームを再マウントするだけで、変更を簡単に破棄できます。

仮想ゲスト環境では、物理サーバの場合と同様に、サーバのディスク上に作成するLVM論理ボリュームに対してスナップショット機能を使用できます。

仮想ホスト環境では、スナップショット機能を使用して、仮想マシンのストレージバックエンドをバックアップしたり、仮想マシンイメージに対する変更(パッチやアップグレードなど)を、ソース論理ボリュームを変更せずにテストしたりできます。仮想マシンは、仮想ディスクファイルの使用ではなく、ストレージバックエンドとして、LVM論理ボリュームを使用する必要があります。LVM論理ボリュームをマウントし、ファイルに格納されたディスクとして仮想マシンイメージを保存するために使用できます。また、そのLVM論理ボリュームを物理ディスクとして割り当てて、ブロックデバイスとして書き込むことができます。

LVM論理ボリュームスナップショットは、シンプロビジョニングできます。サイズを指定しないでスナップショットを作成した場合は、シンプロビジョニングと想定されます。スナップショットは、シンプルから必要な領域を使用するシンボリュームとして作成されます。シンスナップショットボリュームは、他のシンボリュームと同じ特性を持ちます。ボリュームは個別にアクティブ化、拡張、名前変更、および削除でき、そのスナップショットを作成することもできます。

! 重要: クラスタにおけるシンプロビジョニングボリューム

クラスタでシンプロビジョニングスナップショットを使用するには、ソース論理ボリュームとそのスナップショットを1つのクラスタリソースで管理する必要があります。これにより、ボリュームとそのスナップショットを常に同じノードに排他的にマウントできます。

スナップショットが不要になったら、必ず、システムからスナップショットを削除してください。ソース論理ボリュームでデータブロックが変化していくのに応じて、スナップショットは最終的に満杯になります。スナップショットは満杯になると使用不可になるので、ソース論理ボリュームの再マウントができなくなります。

ソース論理ボリュームのスナップショットを複数作成している場合、スナップショットの削除は、最後に作成したものを最初に削除するという順番で行います。

6.2 LVMによるLinuxスナップショットの作成

LVM (Logical Volume Manager)は、ファイルシステムのスナップショットの作成に使用できます。

端末を開いて、次のコマンドを入力します。

```
> sudo lvcreate -s [-L <size>] -n SNAP_VOLUME SOURCE_VOLUME_PATH
```

サイズを指定しない場合、スナップショットはシンスナップショットとして作成されます。

例:

```
> sudo lvcreate -s -L 1G -n linux01-snap /dev/lvm/linux01
```

スナップショットが/dev/lvm/linux01-snapボリュームとして作成されます。

6.3 スナップショットの監視

端末を開いて、次のコマンドを入力します。

```
> sudo lvdisplay SNAP_VOLUME
```

例:

```
> sudo lvdisplay /dev/vg01/linux01-snap

--- Logical volume ---
  LV Name          /dev/lvm/linux01
  VG Name          vg01
  LV UUID          QHVJYh-PR3s-A4SG-s4Aa-MyWN-Ra7a-HL47KL
  LV Write Access  read/write
  LV snapshot status active destination for /dev/lvm/linux01
  LV Status        available
  # open           0
  LV Size          80.00 GB
  Current LE       1024
  COW-table size   8.00 GB
  COW-table LE     512
  Allocated to snapshot 30%
  Snapshot chunk size 8.00 KB
  Segments         1
  Allocation       inherit
  Read ahead sectors 0
```

6.4 Linuxスナップショットの削除

端末を開いて、次のコマンドを入力します。

```
> sudo lvremove SNAP_VOLUME_PATH
```

例:

```
> sudo lvremove /dev/lvmvg/linux01-snap
```

6.5 仮想ホスト上の仮想マシンに対するスナップショットの使用

仮想マシンのバックエンドストレージにLVM論理ボリュームを使用すると、基礎となるデバイスを柔軟に管理でき、ストレージオブジェクトの移動、スナップショットの作成、データのバックアップなどの操作を容易に行うことができるようになります。LVM論理ボリュームをマウントし、ファイルに格納されたディスクとして仮想マシンイメージを保存するために使用できます。また、そのLVM論理ボリュームを物理ディスクとして割り当てて、ブロックデバイスとして書き込むことができます。LVM論理ボリューム上に仮想ディスクイメージを作成してから、LVMのスナップショットを作成できます。

スナップショットの読み込み/書き込み機能を利用して、1つの仮想マシンのインスタンスを複数作成できます。この場合、変更は、仮想マシンの特定のインスタンスのスナップショットに対して行われます。LVM論理ボリューム上に仮想ディスクイメージを作成してソース論理ボリュームのスナップショットを作成し、仮想マシンの特定のインスタンスのスナップショットを変更できます。ソース論理ボリュームのスナップショットをもう1つ作成して、仮想マシンの別のインスタンス用に変更できます。複数の仮想マシンインスタンスのデータの大部分は、ソース論理ボリューム上のイメージと共に存在します。

スナップショットの読み込み/書き込み機能を利用すると、仮想ディスクイメージを維持したまま、ゲスト環境でパッチやアップグレードをテストすることもできます。そのイメージが含まれるLVMボリュームのスナップショットを作成し、そのスナップショットの場所で仮想マシンを実行します。ソース論理ボリュームは変更されず、そのマシンに対する変更はすべてスナップショットに書き込まれます。仮想マシンイメージのソース論理ボリュームに戻るには、仮想マシンの電源をオフにした後、ソース論理ボリュームからスナップショットを削除します。もう一度やり直すには、スナップショットを再作成してマウントしてから、スナップショットイメージ上で仮想マシンを再起動します。

次の手順では、ファイルに格納された仮想ディスクイメージとXenハイパーバイザを使用します。本項の手順は、KVMなど、SUSE Linux Enterpriseプラットフォーム上で動作する他のハイパーバイザに適用できます。スナップショットボリュームからファイルに格納された仮想マシンイメージを実行するには:

1. ファイルに格納された仮想マシンイメージが含まれるソース論理ボリュームがマウントされていることを確認します(たとえば、マウントポイント`/var/lib/xen/images/<IMAGE_NAME>`)。
2. 予想される差分を保存するのに十分な領域があるLVM論理ボリュームのスナップショットを作成します。

```
> sudo lvcreate -s -L 20G -n myvm-snap /dev/lvmmvg/myvm
```

サイズを指定しない場合、スナップショットはシンスナップショットとして作成されます。

3. スナップショットボリュームをマウントするマウントポイントを作成します。

```
> sudo mkdir -p /mnt/xen/vm/myvm-snap
```

4. 作成したマウントポイントにスナップショットボリュームをマウントします。

```
> sudo mount -t auto /dev/lvmmvg/myvm-snap /mnt/xen/vm/myvm-snap
```

5. テキストエディタで、ソース仮想マシンの設定ファイルをコピーし、マウントしたスナップショットボリューム上の、ファイルに格納されたイメージファイルを指すようにパスを変更し、ファイルを`/etc/xen/myvm-snap.cfg`などの名前で保存します。
6. 仮想マシンのマウント済みスナップショットボリュームを使用して、仮想マシンを起動します。

```
> sudo xm create -c /etc/xen/myvm-snap.cfg
```

7. (オプション)スナップショットを削除して、ソース論理ボリューム上の変更されていない仮想マシンイメージを使用します。

```
> sudo umount /mnt/xenvms/myvm-snap  
> sudo lvremove -f /dev/lvmmvg/mylvm-snap
```

8. (オプション)このプロセスを必要なだけ繰り返します。

6.6 スナップショットをソース論理ボリュームとマージして変更を元に戻すか、前の状態にロールバックする

スナップショットは、ボリューム上のデータを前の状態にロールバックまたは復元する必要がある場合に役立ちます。たとえば、管理者の手違いがあった場合、またはパッケージのインストールやアップグレードが失敗したり望む内容と違ったりした場合、データ変更を元に戻さなければならないことがあります。

`lvconvert --merge`コマンドを使用して、LVM論理ボリュームの変更を元に戻すことができます。マージは次のように開始されます。

- ソース論理ボリュームとスナップショットボリュームが両方とも開かれていない場合、マージはすぐに開始されます。
- ソース論理ボリュームまたはスナップショットボリュームが開かれていない場合、初めてソース論理ボリュームまたはスナップショットボリュームのどちらかがアクティブになって両方が閉じられた時点でマージが開始されます。
- ルートファイルシステムのように、閉じることができないソース論理ボリュームの場合、次にサーバが再起動されてソース論理ボリュームがアクティブになるときまで、マージは延期されます。
- ソース論理ボリュームに仮想マシンイメージが含まれる場合、仮想マシンをシャットダウンしてソース論理ボリュームとスナップショットボリュームを(この順序でマウント解除することによって)非アクティブにした後、`merge`コマンドを発行する必要があります。マージ完了時に、ソース論理ボリュームが自動的に再マウントされてスナップショットボリュームは削除されるので、マージ完了後まで仮想マシンを再起動しないでください。マージが完了した後、生成された論理ボリュームを仮想マシンで使用します。

マージが開始されると、サーバ再起動後もマージは自動的に続行され、これはマージが完了するまで続きます。マージの進行中は、マージ中のソース論理ボリュームの新しいスナップショットを作成することはできません。

マージの進行中は、ソース論理ボリュームに対する読み込みまたは書き込みは、マージ中のスナップショットに透過的にリダイレクトされます。これにより、ユーザーは直接、スナップショット作成時のデータを表示したり、そのデータにアクセスしたりできます。マージが完了するまで待つ必要はありません。

マージが完了すると、ソース論理ボリュームにはスナップショット作成時と同じデータと、マージ開始後に行われたデータ変更がすべて含まれます。生成された論理ボリュームの名前、マイナー番号、およびUUIDは、ソース論理ボリュームと同じです。ソース論理ボリュームは自動的に再マウントされ、スナップショットボリュームは削除されます。

1. 端末を開いて、次のコマンドを入力します。

```
> sudo lvconvert --merge [-b] [-i SECONDS] [SNAP_VOLUME_PATH[...snapN]|@VOLUME_TAG]
```

コマンドラインで1つ以上のスナップショットを指定できます。または、複数のソース論理ボリュームに同じボリュームタグを設定し、「@<VOLUME_TAG>」と指定することもできます。タグ付きボリュームのスナップショットは、それぞれのソース論理ボリュームにマージされます。タグ付き論理ボリュームについては、[5.10項 「LVM2ストレージオブジェクトへのタグ付け」](#)を参照してください。

次のオプションがあります。

-b,

--background

デーモンをバックグラウンドで実行します。これにより、指定した複数のスナップショットのマージを同時に並行して実行できます。

-i,

--interval <SECONDS>

進行状況を定期的にパーセント値でレポートします。間隔を秒単位で指定します。

このコマンドの詳細については、[lvconvert\(8\)](#)のマニュアルページを参照してください。

例:

```
> sudo lvconvert --merge /dev/lvmvg/linux01-snap
```

このコマンドは、[/dev/lvmvg/linux01-snap](#)をそのソース論理ボリュームにマージします。

```
> sudo lvconvert --merge @mytag
```

[lvol1](#)、[lvol2](#)、および[lvol3](#)がすべて[mytag](#)でタグ付けされている場合、各スナップショットボリュームは、それぞれのソース論理ボリュームと順番につまり[lvol1](#)、[lvol2](#)、[lvol3](#)の順にマージされます。[--background](#)オプションが指定されている場合、各タグ付きソース論理ボリュームのスナップショットは同時に並行してマージされます。

2. (オプション)ソース論理ボリュームとスナップショットが両方とも開いていて、閉じることができる場合、手動でソース論理ボリュームを非アクティブにしてからアクティブにすることによって、すぐにマージを開始できます。

```
> sudo umount ORIGINAL_VOLUME  
> sudo lvchange -an ORIGINAL_VOLUME  
> sudo lvchange -ay ORIGINAL_VOLUME  
> sudo mount ORIGINAL_VOLUME MOUNT_POINT
```

例:

```
> sudo umount /dev/lvmvg/lvol01  
> sudo lvchange -an /dev/lvmvg/lvol01  
> sudo lvchange -ay /dev/lvmvg/lvol01  
> sudo mount /dev/lvmvg/lvol01 /mnt/lvol01
```

3. (オプション)ソース論理ボリュームとスナップショットボリュームが両方とも開いていて、ソース論理ボリュームを閉じることができない場合(rootファイルシステムなど)、サーバを再起動してソース論理ボリュームをマウントすることによって、再起動後すぐにマージを開始できます。

III ソフトウェアRAID

- 7 ソフトウェアRAIDの設定 98
- 8 ルートパーティション用のソフトウェアRAIDの設定 106
- 9 ソフトウェアRAID 10デバイスの作成 113
- 10 ディグレードRAIDアレイの作成 128
- 11 mdadmによるソフトウェアRAIDアレイのサイズ変更 130
- 12 MDソフトウェアRAID用のストレージエンクロージャLEDユーティリティ 139
- 13 ソフトウェアRAIDのトラブルシューティング 147

7 ソフトウェアRAIDの設定

RAID (Redundant Array of Independent Disks)の目的は、複数のハードディスクパーティションを1つの大きい仮想ハードディスクに結合し、パフォーマンスとデータのセキュリティを最適化することです。ほとんどのRAIDコントローラはSCSIプロトコルを使用します。これは、IDEプロトコルよりも効率的な方法で多数のハードディスクのアドレスを指定でき、コマンドのパラレル処理に適しているからです。一方、IDEハードディスクやSATAハードディスクをサポートしているRAIDコントローラもあります。ソフトウェアRAIDは、ハードウェアRAIDコントローラ購入による追加コストなしで、RAIDシステムの利点を提供します。ただし、これにはいくらかのCPU時間を要し、高性能なコンピュータには適さないメモリ要件があります。

! 重要: クラスタファイルシステムのRAID

クラスタファイルシステムのソフトウェアRAIDはクラスタマルチデバイス(Cluster MD)を使用して設定する必要があります。を参照してください。 Administration Guide for SUSE Linux Enterprise High Availability (<https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/cha-ha-cluster-md.html>) ↗

SUSE Linux Enterpriseには、いくつかのハードディスクを1つのソフトウェアRAIDシステムに統合するオプションがあります。RAIDには、それぞれが異なる目標、利点、および属性をもついくつかのハードディスクを1つのRAIDシステムに結合するためのいくつかの戦略が含まれています。これらは通常、**RAIDレベル**と呼ばれます。

7.1 RAIDレベルの理解

本項では、通常のRAIDレベル(0、1、2、3、4、5)とネストしたRAIDレベルについて説明します。

7.1.1 RAID 0

このレベルでは、各ファイルのブロックが複数のディスクに分散されるので、データアクセスのパフォーマンスが向上します。このレベルはデータのバックアップを提供しないため、実際にはRAIDではありませんが、この種のシステムでは「RAID 0」という名前が一般的です。RAID 0では、2つ以上のハードディスクが互いにプールします。高いパフォーマンスが得られます。ただし、1つのハードディスクに障害が発生しただけで、RAIDシステムが破壊され、データは失われます。

7.1.2 RAID 1

このレベルは、データが別のハードディスク1.1にコピーされるため、データに十分なセキュリティを提供します。これは「ハードディスクミラーリング」と呼ばれます。ディスクが破壊された場合は、ディスクの内容のコピーをミラー先のもう1つのディスクで利用できます。したがって、1つのディスク以外のすべてのディスクが損なわっても、データを保全できます。ただし、損傷が検出されない場合は、正しいディスクに損傷したデータがミラーリングされる可能性があり、その場合はデータが壊れます。単一ディスクアクセスの使用時と比較すると、コピープロセスで書き込みのパフォーマンスが若干低下しますが(10~20%遅くなる)、読み取りアクセスは、通常の物理ハードディスクのどれと比べても、著しく高速です。これは、データが複製されているので、それらを並行してスキャンできるためです。RAID 1では、一般に、読み取りトランザクションの速度が単一ディスクのほぼ2倍、書き込みトランザクションの速度が単一ディスクと同じです。

7.1.3 RAID 2およびRAID 3

これらは、一般的なRAID実装ではありません。レベル2では、データは、ブロックレベルではなく、ビットレベルでストライプ化されます。レベル3は、専用パリティディスクによってバイトレベルのストライプ化を提供しますが、複数の要求を同時にサービスすることはできません。両レベルとも、まれにしか使用されません。

7.1.4 RAID 4

レベル4は、専用パリティディスクと結合されたレベル0と同様に、ブロックレベルのストライピングを提供します。データディスクがエラーになると、パリティデータで置き換え用のディスクが作成されます。ただし、書き込みアクセスでパリティディスクがボトルネックになることがあります。にもかかわらず、レベル4は時々使用されます。

7.1.5 RAID 5

RAID 5は、レベル0とレベル1の間をパフォーマンスおよび冗長性の面で調整して、最適化したものです。ハードディスクスペースは、使用する合計ディスク台数から1を減算した台数分に等しくなります。データは、RAID 0の場合と同様に、ハードディスク間に配布されます。パーティションの1つで作成される「パリティブロック」は、セキュリティ上の理由で存在します。各パーティションはXORによって互いにリンクされているので、システム障害が発生すると、対応するパリティブロックによってディスクの内容が再構築されます。RAID 5の場

合、同時に複数のハードディスクが障害を起こすことはありません。1つのハードディスクに障害がある場合は、可能であればそのハードディスクを交換して、データ消失の危険性をなくす必要があります。

7.1.6 RAID 6

RAID 6は、RAID 5の拡張であり、2つ目の独立した分散パリティスキーム(デュアルパリティ)の使用により、耐障害性をさらに追加します。データ回復プロセスで、2つのハードディスクに障害が発生しても、システムは稼動し続け、データが失われることはありません。

RAID 6は、複数の同時ドライブエラーに耐えることで、非常に高いデータ耐障害性を提供します。RAID 6は、データを失うことなく、2つのデバイスの喪失を処理します。したがって、N個のドライブのデータを保存するには、N+2個のドライブが必要です。その結果、最低限4個のデバイスが必要となります。

通常モードおよび單一ディスク障害モードでは、RAID 5と比べ、RAID 6のパフォーマンスは若干低いですが、同程度です。2台のディスクに障害が発生するとRAID 6はきわめて低速になります。RAID 6設定では、書き込み操作のためにかなりのCPU時間とメモリが必要です。

表 7.1: RAID 5とRAID 6の比較

機能	RAID 5	RAID 6
デバイス数	N+1(最小限3個)	N+2(最小限4個)
パリティ	分散型、シングル	分散型、デュアル
パフォーマンス	書き込みおよび再構築に中程度の影響	シーケンシャルな書き込みでは、RAID 5より影響大
耐障害性	1つのコンポーネントデバイスの障害	2つのコンポーネントデバイスの障害

7.1.7 ネストしたコンプレックスRAIDレベル

他にもRAIDレベルが開発されています(RAIDn、RAID 10、RAID 0+1、RAID 30、RAID 50など)。これらの一部は、ハードウェアベンダーによって作成された専有インプリメンテーションです。RAID 10設定の作成例については、[第9章 「ソフトウェアRAID 10デバイスの作成」](#)を参照してください。

7.2 YaSTによるソフトウェアRAID設定

YaSTソフトRAID設定には、YaST Expert Partitionerからアクセスできます。このパーティション設定ツールを使用すると、既存のパーティションを編集および削除したり、ソフトウェアRAIDで使用する新規パーティションを作成したりすることもできます。これらの方針は、RAIDレベル0、1、5、および6の設定に適用されます。RAID 10の設定については、[第9章「ソフトウェアRAID 10デバイスの作成」](#)で説明されています。

1. YaSTを起動してパーティショナを開きます。
2. 必要に応じて、RAID設定で使用するパーティションを作成します。パーティションをフォーマットしたり、パーティションタイプを0xFD Linux RAIDに設定したりしないでください。既存のパーティションを使用する場合、パーティションタイプを変更する必要はありません。YaSTによって自動的に変更されます。詳細については、『展開ガイド』、[第11章「エキスパートパーティショナ」](#)、11.1項「熟練者向けパーティション設定の使用」を参照してください。

ハードディスクのどれかに障害が発生した場合にデータを失うリスクを減らすため(RAID 1、RAID 5)、およびRAID 0のパフォーマンスを最適化するため、異なるハードディスクに保存されているパーティションを使用することを強くお勧めします。

RAID 0の場合は、少なくとも2つのパーティションが必要です。RAID 1に必要なパーティションは2つだけですが、RAID 5の場合は少なくとも3つのパーティションが必要です。RAID 6セットアップでは、少なくとも4つのパーティションが必要です。各パーティションのサイズを同じにすることをお勧めします。パーティションサイズが異なると、各セグメントが提供するサイズは最小のパーティションサイズにとどまるからです。
3. 左のパネルで、RAIDを選択します。
既存のRAID設定のリストが右のパネルに表示されます。
4. [RAID] ページの左下で、RAIDの追加をクリックします。
5. RAID種類を選択し、追加をクリックして、使用可能なデバイスダイアログから適切な数のパーティションを追加します。
オプションで、RAID名でRAIDに名前を割り当てることができます。この名前は、`/dev/md/NAME`として利用可能になります。詳細については、[7.2.1項「RAIDの名前」](#)を参照してください。



図 7.1: RAID 5設定の例

次へで続行します。

6. チャンクサイズを選択し、該当する場合はパリティアルゴリズムを選択します。最適なチャンクサイズは、データのタイプとRAIDのタイプによって変わります。詳細については、https://raid.wiki.kernel.org/index.php/RAID_setup#Chunk_sizesを参照してください。パリティアルゴリズムの詳細については、--layoutオプションの検索時にman 8 mdadmを使用して参照できます。わからない場合は、デフォルト値を使用してください。
7. 役割でボリュームの役割を選択します。ここで選択した内容は、次のダイアログのデフォルト値にのみ影響します。値は次の手順で変更可能です。わからない場合は、RAWボリューム(未フォーマット)を選択します。
8. フォーマットオプションで、パーティションをフォーマットするを選択し、ファイルシステムを選択します。オプションメニューの内容は、ファイルシステムによって異なります。通常は、デフォルト値を変更する必要はありません。
マウントのオプションの下で、パーティションをマウントするを選択してから、マウントポイントを選択します。Fstabオプションをクリックして、このボリュームの特別なマウントオプションを追加します。
9. 完了をクリックします。

10. 次へをクリックし、変更が一覧されることを確認してから、完了をクリックします。

! 重要: ディスク上のRAID

パーティショナはパーティションの代わりにディスクの上にRAIDを作成することを可能にしますが、いくつかの理由のため、このアプローチは推奨されません。このようなRAIDにブートローダをインストールすることはサポートされていないため、ブート用に別のデバイスを使用する必要があります。fdiskやpartedなどのツールは当該RAIDでは適切に機能しないため、RAIDの特定のセットアップを知らない人によって誤った診断やアクションが行われる可能性があります。

7.2.1 RAIDの名前

デフォルトでは、ソフトウェアRAIDデバイスには、mdN(Nは数字)というパターンに従った数字の名前が付いています。そのため、たとえば、/dev/md127としてアクセスでき、/proc/mdstatと/proc/partitionsではmd127としてリストされます。このような名前では作業しづらい場合があります。SUSE Linux Enterprise Serverでは、この問題を回避する方法を2つ提供しています。

デバイスへの名前付きリンクを指定する

オプションで、YaSTでRAIDデバイスを作成する際、またはmdadm --create '/dev/md/NAME'を使用してコマンドラインで、RAIDデバイスの名前を指定できます。デバイス名はmdNのままであるが、リンク/dev/md/NAMEが作成されます。

```
> ls -og /dev/md
total 0
lrwxrwxrwx 1 8 Dec  9 15:11 myRAID -> ../md127
```

デバイスは/procには引き続きmd127としてリストされます。

名前付きデバイスを指定する

ご使用のセットアップでデバイスへの名前付きリンクでは不十分な場合、次のコマンドを実行して、/etc/mdadm.confにCREATE names=yesという行を追加します。

```
> echo "CREATE names=yes" | sudo tee -a /etc/mdadm.conf
```

これにより、myRAIDのような名前が「実際の」デバイス名として使用されるようになります。このデバイスは/dev/myRAIDでアクセスできるだけでなく、/procにもmyRAIDとしてリストされます。これは、設定ファイルの変更後に設定したRAIDにのみ適用される点に注意してください。アクティブなRAIDでは、停止して再アセンブルするまで引き続きmdN形式の名前が使用されます。



警告: 非互換のツール

一部のツールは、名前付きRAIDデバイスをサポートしていません。ツールがRAIDデバイスに`mdN`形式の名前が付いていることを予期している場合、そのツールはデバイスを特定できません。

7.3 AArch64のRAID 5のストライプサイズの設定

デフォルトでは、ストライプサイズは4KBに設定されています。デフォルトのストライプサイズを変更する必要がある場合、たとえば、AArch64の一般的なページサイズの64KBに合わせるには、CLIを使用してストライプサイズを手動で設定できます。

```
> sudo echo 16384 > /sys/block/	md1/md/stripe_size
```

上記のコマンドを実行すると、ストライプサイズは16KBに設定されます。4096、8192など他の値を設定できますが、値は2のべき乗である必要があります。

7.4 ソフトウェアRAIDの監視

`monitor`モードでデーモンとして`mdadm`を実行し、ソフトウェアRAIDを監視することができます。`monitor`モードでは、`mdadm`はアレイのディスク障害を定期的に確認します。障害が発生した場合、`mdadm`は管理者に電子メールを送信します。チェックの時間間隔を定義するには、次のコマンドを実行します。

```
mdadm --monitor --mail=root@localhost --delay=1800 /dev/md2
```

先に示したコマンドは1800秒間隔で`/dev/md2`アレイの監視をオンにします。障害が発生した場合、電子メールが`root@localhost`に送信されます。



注記: デフォルトでは、RAIDチェックが有効化されています

デフォルトでは、RAIDチェックが有効化されています。各チェックの間隔が十分な長さではなく、警告が表示される場合があります。このように、`delay`オプションにより高い値を設定することにより、間隔を増やすことができます。

7.5 詳細情報

ソフトウェアRAIDの設定方法と詳細情報が、次のHOWTOにあります。

- The Linux RAID wiki: <https://raid.wiki.kernel.org/>
- The Software RAID HOWTO(/usr/share/doc/packages/mdadm/Software-RAID.HOWTO.html ファイル)

「linux-raid」 (<https://marc.info/?l=linux-raid>)などのLinux RAIDメーリングリストもあります。

8 ルートパーティション用のソフトウェアRAIDの設定

SUSE Linux Enterprise Serverでは、Device Mapper RAIDツールがYaSTパーティショナに統合されています。インストール時にパーティショナを使用して、ルート(/)パーティションを含むシステムデバイス用にソフトウェアRAIDを作成することができます。`/boot`パーティションは、RAID 1以外のRAIDパーティションには保存できません。

! 重要: RAID 1の`/boot/efi`はブートしない場合がある

RAID上に`/boot/efi`パーティションを作成する際、ファームウェアがRAID上のブートパーティションを認識しない場合があることに注意してください。その場合、ファームウェアはブートを拒否します。

8.1 ルートパーティション用のソフトウェアRAIDデバイスを使用するための前提条件

設定が次の要件を満たしていることを確認してください。

- RAID 1のミラーリングデバイスを作成するため、2つのハードドライブが必要です。ハードドライブは類似のサイズで構成する必要があります。RAIDは小さい方のドライブのサイズを採用します。ブロックストレージデバイスには、ローカル(マシンに内蔵、または直結されたもの)、ファイバチャネルストレージサブシステム、またはiSCSIストレージサブシステムを自由に組み合わせることができます。
- ブートローダをMBRにインストールする場合、`/boot`用の別のパーティションは必要ありません。ブートローダをMBRにインストールすることが不可能な場合は、`/boot`が別個のパーティションに存在する必要があります。
- UEFIマシンの場合、専用の`/boot/efi`パーティションを設定する必要があります。これはVFATフォーマットである必要があります。RAID 1デバイスに配置されていれば、`/boot/efi`が存在する物理ディスクに障害が発生した場合にブートの問題を回避できます。
- ハードウェアRAIDデバイスを使用している場合は、その上でソフトウェアRAIDを実行しようとしないでください。

- iSCSIターゲットデバイスをご使用の場合は、RAIDデバイスを作成する前にiSCSIイニシエータサポートを有効にする必要があります。
- ご使用のストレージサブシステムが、ソフトウェアRAIDを使用する予定の直接接続されたローカルデバイス、ファイバチャネルデバイス、またはiSCSIデバイスとサーバの間で複数のI/Oパスを提供している場合は、RAIDデバイスを作成する前に、マルチパスサポートを有効にしなければなりません。

8.2 ルート(/)パーティションにソフトウェアRAIDデバイスを使用するシステムの設定

1. YaSTを使用してインストールを開始し、推奨されたパーティション分割の手順に到達するまで、『展開ガイド』、第9章「インストール手順」の説明に従って進めます。
2. エキスパートパーティショナをクリックして、カスタムパーティショニングツールを開きます。推奨される提案を使用するか、既存の提案を使用することができます。
3. (オプション)使用したいiSCSIターゲットデバイスがある場合、画面左上のセクションでシステム>設定>Configure iSCSI (iSCSIの設定)の順に選択して、iSCSIイニシエータソフトウェアを有効にする必要があります。詳細については、[第15章 「IPネットワークの大容量記憶域: iSCSI」](#)を参照してください。
4. (オプション)使用したいFCoEターゲットデバイスがある場合、画面左上のセクションでシステム>設定>Configure iSCSI (iSCSIの設定)の順にクリックして、インターフェースを設定する必要があります。
5. (オプション)パーティショニングの変更を破棄する必要がある場合は、システム>デバイスの再検出をクリックします。
6. ソフトウェアRAIDに使用する各デバイスのLinux RAIDフォーマットを設定します。/、/boot/efi、またはスワップパーティションにはRAIDを使用する必要があります。
 - a. 左パネルでハードディスクを選択し、使用するデバイスを選択してからパーティションの追加をクリックします。
 - b. 新しいパーティションのサイズで、使用するサイズを指定し、次に次へをクリックします。
 - c. 役割でRaw Volume (Unformatted) (RAWボリューム(未フォーマット))を選択します。

- d. Do not format (フォーマットしない)とDo not mount (マウントしない)を選択し、パーティションIDをLinux RAIDに設定します。
- e. 次へをクリックし、2番目のパーティションに対して同じ手順を繰り返します。



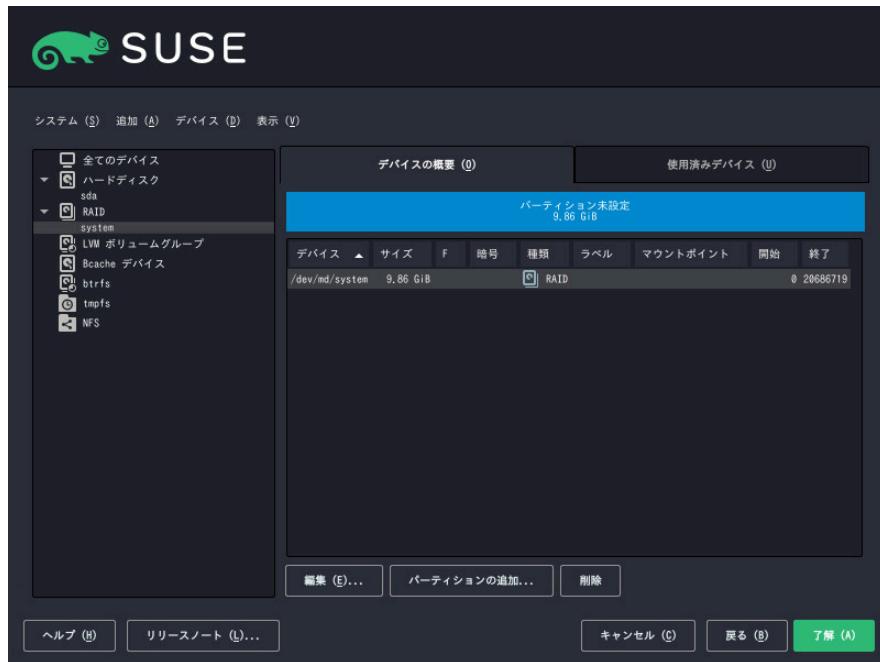
7. /パーティション用のRAIDデバイスを作成します。

- a. 左パネルでRAIDを選択し、RAIDの追加を選択します。
- b. /パーティションに対して目的のRAIDの種類を設定し、RAID名をsystemに設定します。
- c. 前の手順で準備した2つのRAIDデバイスを使用可能なデバイスから選択し、追加をクリックして追加します。

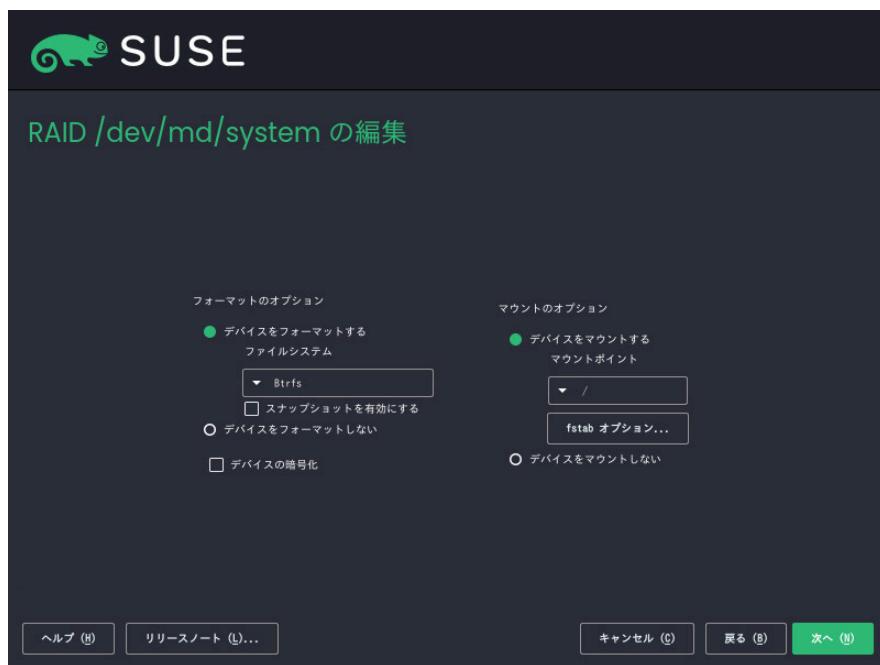


次へで続行します。

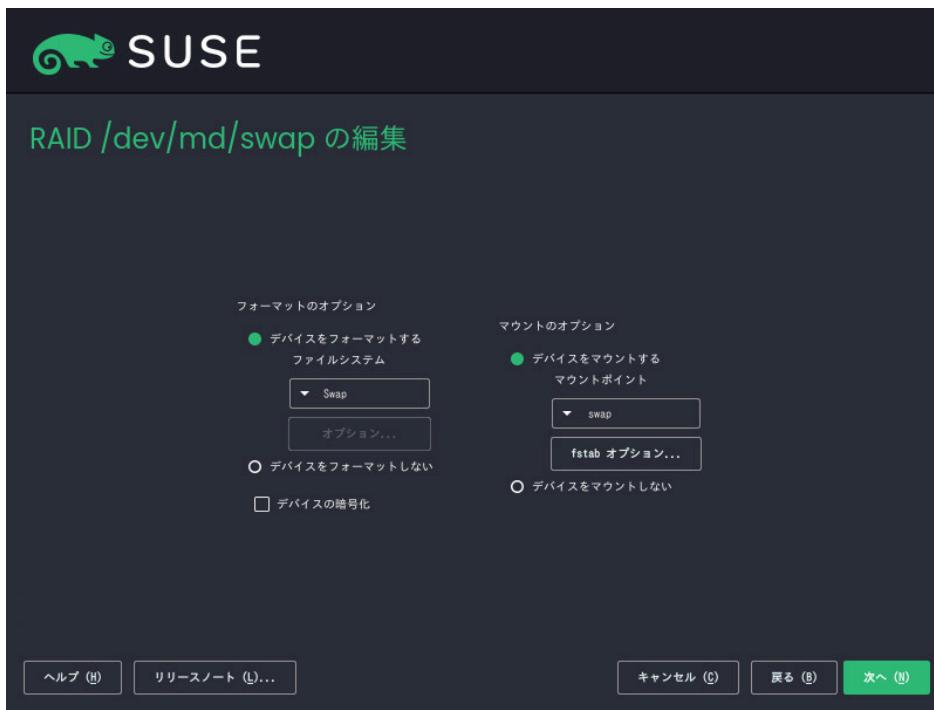
- d. ドロップダウンボックスからチャンクサイズを選択します。デフォルト値をそのまま使用するのが安全です。
- e. 左のパネルで、RAIDをクリックします。デバイスの概要タブで、新しいRAIDを選択し、編集をクリックします。



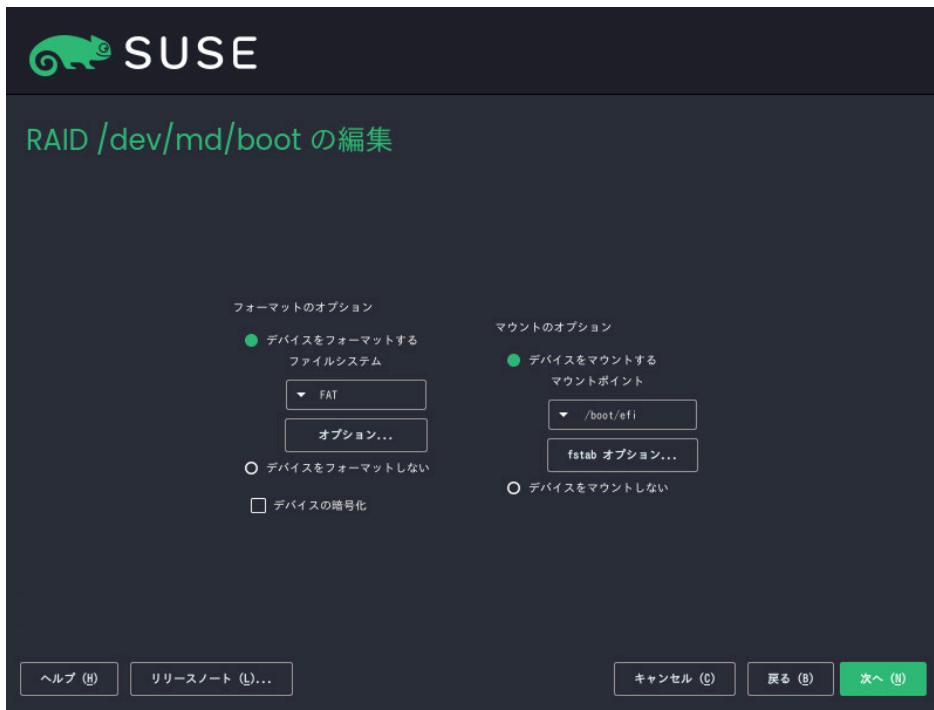
- f. 役割で、オペレーティングシステムを選択し、次へで続行します。
- g. ファイルシステムを選択し、マウントポイントを`/`に設定します。Nextをクリックして、ダイアログを終了します。



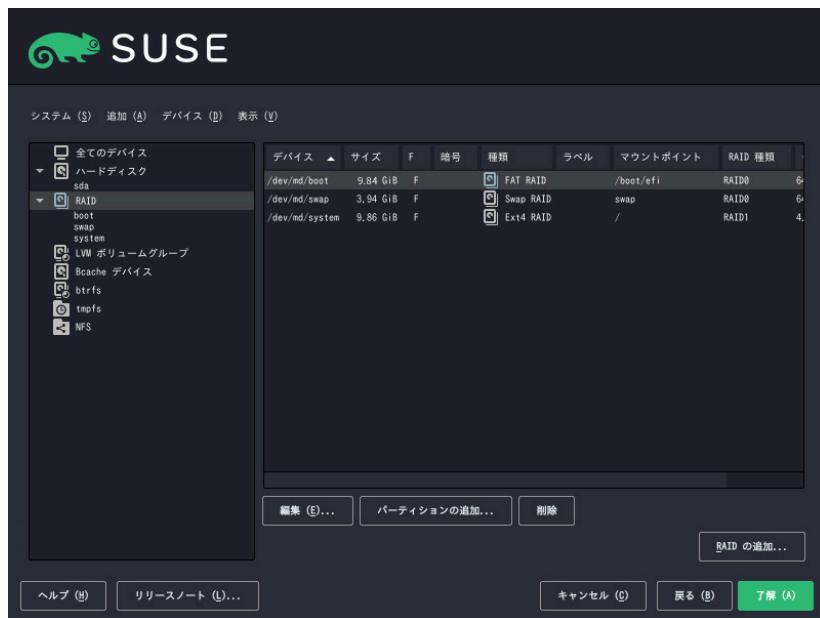
8. ソフトウェアRAIDデバイスはデバイスマッパーによって管理され、デバイスを`/dev/md/system`パスの下に作成します。
9. オプションで、RAIDでスワップパーティションを作成できます。先に説明された手順と同様の手順を使いますが、役割からスワップを選択します。ファイルシステムとマウントポイントを以下のように選択します。次へをクリックします。



10. オプションで、UEFIマシンに対し、同様の手順を使用して、/boot/efiにマウントされるパーティションを作成することもできます。/boot/efiでサポートされるのはRAID 1のみであること、およびパーティションがFAT32ファイルシステムでフォーマットされている必要があることを忘れないでください。



パーティショニングは次のようになります。



11. 承認をクリックして、パーティショナを終了します。

推奨されたパーティション分割ページに新しい案が表示されます。

12. インストールを続行します。独立した`/boot/efi`パーティションを持つUEFIマシンでは、インストールの設定画面で起動をクリックし、EFI向けGRUB2をブートローダとして設定します。Secure Bootサポートを有効にするオプションがアクティブになっていることを確認します。

サーバを再起動するたびに、デバイスマッパーが起動時に開始し、ソフトウェアRAIDが自動的に認識され、ルート(/)パーティション上のオペレーティングシステムを開始することができます。

9 ソフトウェアRAID 10デバイスの作成

本項では、ネストしたコンプレックスRAID 10デバイスの設定方法について説明します。RAID 10デバイスは、ネストしたRAID 1(ミラーリング)アレイとRAID 0(ストライピング)アレイで構成されます。ネストしたRAIDは、ストライピングミラー(RAID 1+0)またはミラーリングされたストライプ(RAID 0+1)のいずれかとして設定できます。コンプレックスRAID 10のセットアップは、ミラーとストライプを組み合わせ、より高いデータ冗長性レベルをサポートすることによってデータのセキュリティを強化します。

9.1 mdadmによるネストしたRAID 10デバイスの作成

ネストしたRAIDデバイスは、物理ディスクを使用する代わりに、その基本エレメントとして別のRAIDアレイを使用するRAIDアレイで構成されます。この構成の目的は、RAIDのパフォーマンスと耐障害性を向上することです。ネストしたRAIDレベルの設定はYaSTではサポートされていませんが、mdadmコマンドラインツールを使用して実行できます。

ネストの順序に基づいて、2つの異なるネストしたRAIDを設定できます。このマニュアルでは、次の用語を使用します。

- **RAID 1+0:** まず、RAID 1(ミラー)アレイが構築され、次に、それらのアレイが組み合わされてRAID 0(ストライプ)アレイを構成します。
- **RAID 0+1:** まず、RAID 0(ストライプ)アレイが構築され、次に、それらのアレイが組み合わされてRAID 1(ミラー)アレイを構成します。

次の表では、RAID 10ネスティングの欠点と利点を、1+0対0+1という形式で説明します。使用的ストレージオブジェクトは、それぞれが専用のI/Oをもつ別々のディスクに常駐すると想定しています。

表 9.1: ネストしたRAIDレベル

RAIDレベル	説明	パフォーマンスと耐障害性
10 (1+0)	RAID 1(ミラー)アレイで構築されたRAID (ストライプ)	RAID 1+0は、高レベルのI/Oパフォーマンス、データ冗長性、およびディスク耐障害性を提供します。RAIDの各メンバーデバイスは個々にミラーリングされるので、エラーになったディスクのミラー先が異なる限り、複数ディスクの障害を許容し、データを使用することができます。

RAIDレベル	説明	パフォーマンスと耐障害性
		オプションとして、ベースをなすミラーリングされたアレイごとにスペアを設定したり、すべてのミラーに対するスペアグループに対応するスペアを設定できます。
10 (0+1)	RAID 0(ストライプ)アレイで構築されたRAID 1(ミラー)	<p>RAID 0+1は、高レベルのI/Oパフォーマンスとデータ冗長性を提供しますが、耐障害性が1+0よりも若干低くなります。ミラーの一方のサイドで複数のディスクがエラーになると、もう一方のミラーが使用可能になります。ただし、ミラーの両サイドで同時にディスクが失われると、すべてのデータが喪失します。</p> <p>このソリューションは1+0ソリューションより耐障害性が低いですが、別のサイトで保守を実行したり、ミラーを保持する必要がある場合、ミラーのサイド全体をオフラインにしても、完全に機能するストレージデバイスを保持することができます。また、2つのサイト間の接続が失われた場合は、どちらかのサイトがもう一方のサイトから独立して稼動します。ミラーリングされたセグメントをストライプする場合はこうなりません。ミラーが低レベルで管理されているからです。</p> <p>デバイスがエラーになると、RAID 0には耐障害性がないので、そのサイドのミラーがエラーになります。新しいRAID 0を作成して、エラーになったサイドに置き換え、次に、ミラーを再同期してください。</p>

9.1.1 mdadmによるネストしたRAID 10 (1+0)デバイスの作成

ネストしたRAID 1+0は、2つ以上のRAID 1(ミラー)デバイスを作成し、それらのRAID 1デバイスをRAID 0のコンポーネントデバイスとして使用することで構築します。



重要: マルチパス処理

デバイスに対する複数の接続を管理する必要がある場合は、マルチパスI/Oを設定してから、RAIDデバイスを設定する必要があります。詳細については、[第18章 「デバイスのマルチパスI/Oの管理」](#)を参照してください。

本項の手順では、次の表に示すデバイス名を使用します。それらのデバイス名は、必ず、ご使用のデバイスの名前で変更してください。

表9.2: ネスティングでRAID 10 (1+0)を作成するシナリオ

rawデバイス	RAID 1(ミラー)	RAID 1+0(ストライピングミラー)
/dev/sdb1	/dev/md0	/dev/md2
/dev/sdc1		
/dev/sdd1	/dev/md1	
/dev/sde1		

1. 端末を開きます。
2. 必要に応じて、partedなどのディスクパーティショナを使用して、同じサイズの0xFD Linux RAIDパーティションを4つ作成します。
3. 1デバイスごとに2つの異なるデバイスを使用して、2つのソフトウェアRAID 1デバイスを作成します。コマンドプロンプトで、次の2つのコマンドを入力します。

```
> sudo mdadm --create /dev/md0 --run --level=1 --raid-devices=2 /dev/sdb1 /dev/sdc1  
sudo mdadm --create /dev/md1 --run --level=1 --raid-devices=2 /dev/sdd1 /dev/sde1
```

4. ネストしたRAID 1+0デバイスを作成します。コマンドプロンプトで、前の手順で作成したソフトウェアRAID 1デバイスを使用して、次のコマンドを入力します。

```
> sudo mdadm --create /dev/md2 --run --level=0 --chunk=64 \  
--raid-devices=2 /dev/md0 /dev/md1
```

デフォルトのチャンクサイズは 64KBです。

5. RAID 1+0デバイス/dev/md2上でファイルシステム(XFSファイルシステムなど)を作成します。

```
> sudo mkfs.xfs /dev/md2
```

これとは別のファイルシステムを使用するには、コマンドを変更します。

6. /etc/mdadm.conf ファイルを編集するか、ファイルがまだ存在しない場合は作成します(たとえば、sudo vi /etc/mdadm.conf を実行します)。次の行を追加します(ファイルが既に存在する場合、最初の行は記述済みの可能性があります)。

```
DEVICE containers partitions
ARRAY /dev/md0 UUID=UUID
ARRAY /dev/md1 UUID=UUID
ARRAY /dev/md2 UUID=UUID
```

各デバイスのUUIDは次のコマンドで取得できます。

```
> sudo mdadm -D /dev/DEVICE | grep UUID
```

7. /etc/fstab ファイルを編集して、RAID 1+0デバイス /dev/md2 のエントリを追加します。次の例は、XFSファイルシステム、およびマウントポイントとして/dataを使用する RAIDデバイスのエントリを示しています。

```
/dev/md2 /data xfs defaults 1 2
```

8. RAIDデバイスをマウントします。

```
> sudo mount /data
```

9.1.2 mdadmによるネストしたRAID 10 (0+1)デバイスの作成

ネストしたRAID 0+1は、2個から4個のRAID 0(ストライプ)デバイスで構築され、それらのRAID 0デバイスをミラーリングしてRAID 1のコンポーネントデバイスとします。

! 重要: マルチパス処理

デバイスに対する複数の接続を管理する必要がある場合は、マルチパスI/Oを設定してから、RAIDデバイスを設定する必要があります。詳細については、[第18章 「デバイスのマルチパスI/Oの管理」](#)を参照してください。

この構成では、RAID 0がデバイスの喪失に耐えられないので、ベースのRAID 0デバイスにスペアデバイスを指定できません。デバイスがミラーの1つのサイドでエラーになった場合は、置き換え用のRAID 0デバイスを作成して、ミラーに追加します。

本項の手順では、次の表に示すデバイス名を使用します。それらのデバイス名は、必ず、ご使用のデバイスの名前で変更してください。

表9.3: ネスティングでRAID 10 (0+1)を作成するシナリオ

rawデバイス	RAID 0 (ストライプ)	RAID 0+1 (ミラー化ストライピング)
/dev/sdb1 /dev/sdc1	/dev/md0	/dev/md2
/dev/sdd1 /dev/sde1	/dev/md1	

1. 端末を開きます。
2. 必要に応じて、partedなどのディスクパーティショナを使用して、同じサイズの0xFD Linux RAIDパーティションを4つ作成します。
3. RAID 0デバイスごとに2つの異なるデバイスを使用して、2つのソフトウェアRAID 0デバイスを作成します。コマンドプロンプトで、次の2つのコマンドを入力します。

```
> sudo mdadm --create /dev/md0 --run --level=0 --chunk=64 \
--raid-devices=2 /dev/sdb1 /dev/sdc1
sudo mdadm --create /dev/md1 --run --level=0 --chunk=64 \
--raid-devices=2 /dev/sdd1 /dev/sde1
```

デフォルトのチャンクサイズは 64KBです。

4. ネストしたRAID 0+1デバイスの作成コマンドプロンプトで、前の手順で作成したソフトウェアRAID 0デバイスを使用して、次のコマンドを入力します。

```
> sudo mdadm --create /dev/md2 --run --level=1 --raid-devices=2 /dev/md0 /dev/md1
```

5. RAID 1+0デバイス//dev/md2上でファイルシステム(XFSファイルシステムなど)を作成します。

```
> sudo mkfs.xfs /dev/md2
```

これとは別のファイルシステムを使用するには、コマンドを変更します。

6. /etc/mdadm.confファイルを編集するか、ファイルがまだ存在しない場合は作成します(たとえば、sudo vi /etc/mdadm.confを実行します)。次の行を追加します(ここでも、ファイルが既に存在する場合、最初の行は記述済みの可能性があります)。

```
DEVICE containers partitions
ARRAY /dev/md0 UUID=UUID
ARRAY /dev/md1 UUID=UUID
```

```
ARRAY /dev/md2 UUID=UUID
```

各デバイスのUUIDは次のコマンドで取得できます。

```
> sudo mdadm -D /dev/DEVICE | grep UUID
```

7. `/etc/fstab` ファイルを編集して、RAID 1+0 デバイス `/dev/md2` のエントリを追加します。次の例は、XFS ファイルシステム、およびマウントポイントとして `/data` を使用する RAID デバイスのエントリを示しています。

```
/dev/md2 /data xfs defaults 1 2
```

8. RAID デバイスをマウントします。

```
> sudo mount /data
```

9.2 コンプレックスRAID 10の作成

YaST(および`mdadm`と`--level=10`オプション)では、RAID 0(ストライピング)およびRAID 1(ミラーリング)の両方の機能を組み合わせた単一のコンプレックスソフトウェアRAID 10デバイスが作成されます。すべてのデータブロックの複数のコピーが、ストライピングの規則に従って、複数のドライブ上に配置されます。コンポーネントデバイスは、すべて同じサイズにする必要があります。

コンプレックスRAIDは、ネストしたRAID 10 (1+0)と目的は同じですが、次の点で異なります。

表 9.4: 複雑なRAID 10とネストしたRAID 10の比較

機能	コンプレックスRAID 10	ネストしたRAID 10 (1+0)
デバイス数	偶数個または奇数個のコンポーネントデバイス	偶数個のコンポーネントデバイス
コンポーネントデバイス	単一のRAIDデバイスとして管理されます。	ネストしたRAIDデバイスとして管理されます。
ストライピング	ストライピングは、コンポーネントデバイス上にnearレイアウトまたはfarレイアウトを生じます。	ストライピングは、連続的に、すべてのコンポーネントデバイスをまたぎます。

機能	コンプレックスRAID 10	ネストしたRAID 10 (1+0)
	farレイアウトでは、RAID 1ペアの数でなく、ドライブ数で増減するシーケンシャルな読み込みスループットを提供します。	
データの複数コピー	2からアレイ内のデバイス数まで	ミラーリングされたセグメントごとにコピー
ホットスペアデバイス	単一スペアですべてのコンポーネントデバイスに対応できます。	ベースをなすミラーリングされたアレイごとにスペアを設定したり、すべてのミラーに対応するスペアグループに対するスペアを設定できます。

9.2.1 コンプレックスRAID 10のデバイスおよびレプリカの数

コンプレックスRAID 10アレイの設定時に、データブロックごとに必要なレプリカ数を指定する必要があります。デフォルトのレプリカ数は2ですが、2からアレイ内のデバイス数まで可能です。

少なくとも、指定のレプリカ数と同数のコンポーネントデバイスを使用する必要があります。ただし、RAID 10アレイのコンポーネントデバイス数は各データブロックのレプリカ数の倍数である必要はありません。有効なストレージサイズは、デバイス数をレプリカ数で割った数です。

たとえば、5個のコンポーネントデバイスで作成したアレイに2つのレプリカを指定した場合は、各ブロックのコピーが2つの異なるデバイスに保存されます。したがって、すべてのデータの1コピーの有効なストレージサイズは、5/2(つまり、コンポーネントデバイスのサイズの2.5倍)となります。

9.2.2 レイアウト

コンプレックスRAID 10のセットアップでは、ディスクにデータブロックを配置する方法を定義するレイアウトが3つサポートされています。利用可能なレイアウトは、near(デフォルト)、far、およびoffsetです。各レイアウトはパフォーマンス特性が異なるため、ワークロードに適したレイアウトを選択することが重要です。

9.2.2.1 nearレイアウト

nearレイアウトでは、異なるコンポーネントデバイス上で、データブロックのコピーが互いに接近してストライプされます。つまり、あるデータブロックの複数のコピーが異なるデバイス内で同様にオフセットされます。nearは、RAID 10のデフォルトレイアウトです。たとえば、奇数個のコンポーネントデバイスとデータの2コピーを使用する場合は、一部のコピーが、1チャンク分、デバイス内を前進します。

コンプレックスRAID 10のnearレイアウトは、半数のドライブ上のRAID 0と同様の読み書きパフォーマンスを提供します。

偶数個のディスクと2つのレプリカを使用したnearレイアウト

```
sda1 sdb1 sdc1 sde1
 0   0   1   1
 2   2   3   3
 4   4   5   5
 6   6   7   7
 8   8   9   9
```

奇数個のディスクと2つのレプリカを使用したnearレイアウト

```
sda1 sdb1 sdc1 sde1 sdf1
 0   0   1   1   2
 2   3   3   4   4
 5   5   6   6   7
 7   8   8   9   9
10  10  11  11  12
```

9.2.2.2 farレイアウト

farレイアウトは、すべてのドライブの前半部分にデータをストライプし、次に、2つ目のデータコピーをすべてのドライブの後半部分にストライプして、ブロックのすべてのコピーが異なるドライブに配置されるようにします。値の2つ目のセットは、コンポーネントドライブの中ほどから開始します。

farレイアウトでは、コンプレックスRAID 10の読み込みパフォーマンスは、すべてのドライブを使用したRAID 0と同様ですが、書き込みパフォーマンスは、ドライブヘッドのシーク回数が増えるので、RAID 0よりかなり遅くなります。このレイアウトは、読み込み専用ファイルサーバなどの、読み込み集約型操作に最適です。

RAID 10の書き込み速度は、nearレイアウトを使用しているRAID 1やRAID 10などの他のミラーリングRAIDの種類と同等です。これは、ファイルシステムのエレベータが生の書き込みよりも効率のよい書き込みのスケジュールを行うためです。RAID 10をfarレイアウトで使用する方法は、ミラーリングによる書き込みアプリケーションに適しています。

偶数個のディスクと2つのレプリカを使用したfarレイアウト

```
sda1 sdb1 sdc1 sde1
 0   1   2   3
 4   5   6   7
 .
 .
 3   0   1   2
 7   4   5   6
```

奇数個のディスクと2つのレプリカを使用したfarレイアウト

```
sda1 sdb1 sdc1 sde1 sdf1
 0   1   2   3   4
 5   6   7   8   9
 .
 .
 4   0   1   2   3
 9   5   6   7   8
```

9.2.2.3 offsetレイアウト

offsetレイアウトでは、あるチャンクの複数のコピーが連續したドライブ上で連續したオフセットにレイアウトされるよう、ストライプが複製されます。実際は、それぞれのストライプが複製され、コピーが1つのデバイスでオフセットされます。これにより、適度な大きさのチャンクサイズを使用している場合は、farレイアウトと同様の読み込み特性が得られますが、書き込みのシーク回数は少なくなります。

偶数個のディスクと2つのレプリカを使用したoffsetレイアウト

```
sda1 sdb1 sdc1 sde1
 0   1   2   3
 3   0   1   2
 4   5   6   7
 7   4   5   6
 8   9   10  11
11   8   9   10
```

奇数個のディスクと2つのレプリカを使用したoffsetレイアウト

sda1	sdb1	sdc1	sde1	sdf1
0	1	2	3	4
4	0	1	2	3
5	6	7	8	9
9	5	6	7	8
10	11	12	13	14
14	10	11	12	13

9.2.2.4 YaSTおよびmdadmによるレプリカ数とレイアウトの指定

レプリカ数とレイアウトは、YaSTではパリティアルゴリズム、mdadmでは`--layout`パラメータで指定します。使用できる値は次のとおりです。

nN

nearレイアウトの場合、nを指定し、Nをレプリカ数で置き換えます。レイアウトおよびレプリカ数を設定しない場合、デフォルトでn2が使用されます。

fN

farレイアウトの場合、fを指定し、Nをレプリカ数で置き換えます。

oN

offsetレイアウトの場合、oを指定し、Nをレプリカ数で置き換えます。



注記: レプリカの数

YaSTでは、パリティアルゴリズムパラメータに設定可能なすべての値が自動的に表示されます。

9.2.3 YaSTパーティショナによるコンプレックスRAID 10の作成

1. YaSTを起動してパーティショナを開きます。
2. 必要に応じて、RAID設定で使用するパーティションを作成します。パーティションをフォーマットしたり、パーティションタイプを0xFD Linux RAIDに設定したりしないでください。既存のパーティションを使用する場合、パーティションタイプを変更する必要はありません。YaSTによって自動的に変更されます。詳細については、『展開ガイド』、第11章「エキスパートパーティショナ」、11.1項「熟練者向けパーティション設定の使用」を参照してください。

RAID 10の場合は、少なくとも4つのパーティションが必要です。ハードディスクのどちらに障害が発生した場合にデータを失うリスクを減らすため、異なるハードディスクに保存されているパーティションを使用することを強くお勧めします。各パーティションのサイズを同じにすることをお勧めします。パーティションサイズが異なると、各セグメントが提供するサイズは最小のパーティションサイズにとどまるからです。

3. 左のパネルで、RAIDを選択します。

既存のRAID設定のリストが右のパネルに表示されます。

4. [RAID] ページの左下で、RAIDの追加をクリックします。

5. RAID種類で、RAID 10 (ミラーリングおよびストライピング)を選択します。

オプションで、RAID Name (RAID名)でRAIDに名前を割り当てることができます。この名前は、`/dev/md/NAME`として利用可能になります。詳細については、[7.2.1項 「RAIDの名前」](#) を参照してください。

6. 使用可能なデバイスリストで、希望のパーティションを選択し、次に追加をクリックして、それらを選択したデバイスリストに移動します。



7. (オプション) 分類をクリックして、RAIDアレイ内でのディスクの好みの順番を指定します。

RAID 10など、追加したディスクの順序が重要なRAIDタイプでは、デバイスの使用順序を指定できます。これにより、アレイの半数を特定のディスクサブシステムに配置し、もう半数を別のディスクサブシステムに配置できます。たとえば、1つのディスクサブシステムに障害が発生した場合、システムは2番目のディスクサブシステムから稼働し続けます。

- a. 各ディスクを順番に選択して、Class Xボタンのいずれかをクリックします。ここで、Xは、そのディスクに割り当てる文字です。用意されているクラスはA、B、C、DおよびEですが、多くの場合必要なクラスはそれより少なくなります(たとえばAとBのみ)。このようにして、すべての利用可能なRAIDディスクを割り当てます。
- 複数のデバイスを選択するには、**Ctrl** キーまたは**Shift** キーを押します。選択したデバイスを右クリックして、コンテキストメニューから適切なクラスを選択することもできます。

- b. 次のソートオプションのいずれかを選択して、デバイスの順序を指定します。

Sorted: クラスAのすべてのデバイスを、クラスBのすべてのデバイスより前に、というように並べます。例: AABBCC。

Interleaved: クラスAの最初のデバイス、次にクラスBの最初のデバイス、次にデバイスが割り当てられたすべての後続のクラスの順に、デバイスを並べます。次にクラスAの2番目のデバイス、クラスBの2番目のデバイス、というように続きます。クラスを持たないデバイスはすべて、デバイスリストの最後に並べられます。例: ABCABC。

Pattern File: それぞれが正規表現とクラス名である、複数の行を含む既存のファイルを選択します("sda.* A")。その正規表現に合致するすべてのデバイスが、その行に指定されたクラスに割り当てられます。正規表現はカーネル名(/dev/sda1)、udevパス名(/dev/disk/by-path/pci-0000:00:1f.2-scsi-0:0:0:0-part1)、そしてudev ID (dev/disk/by-id/ata-ST3500418AS_9VMN8X8L-part1)と照合されます。デバイスの名前が、2つ以上の正規表現に合致する場合は、最初に合致したものでクラスが決定されます。

- c. ダイアログの下で、OKをクリックして、順番を確定します。



8. 次へをクリックします。
9. RAIDオプションで、チャンクサイズとパリティアルゴリズムを指定し、次に次へをクリックします。
RAID 10の場合、パリティオプションは、n (near)、f (far)、およびo (offset)です。数字は、必要となる各データブロックのレプリカの数を示します。2がデフォルトの設定です。詳細については、[9.2.2項「レイアウト」](#)を参照してください。
10. ファイルシステムとマウントオプションをRAIDデバイスに追加して、完了をクリックします。
11. 次へをクリックします。
12. 変更する内容を確認して、完了をクリックすると、RAIDが作成されます。

9.2.4 mdadmによるコンプレックスRAID 10の作成

本項の手順では、次の表に示すデバイス名を使用します。それらのデバイス名は、必ず、ご使用のデバイスの名前で変更してください。

表 9.5: MDADMでRAID 10を作成するシナリオ

rawデバイス	RAID 10
/dev/sdf1	/dev/md3

rawデバイス	RAID 10
/dev/sdg1	
/dev/sdh1	
/dev/sdi1	

1. 端末を開きます。
2. 必要に応じて、partedなどのディスクパーティショナを使用して、同じサイズの0xFD Linux RAIDパーティションを少なくとも4つ作成します。
3. 次のコマンドを入力してRAID 10を作成します。

```
> sudo mdadm --create /dev/md3 --run --level=10 --chunk=32 --raid-devices=4 \
/dev/sdf1 /dev/sdg1 /dev/sdh1 /dev/sdi1
```

--raid-devicesの値とパーティションのリストは、ご使用のセットアップに応じて調整してください。

ここに示すコマンドでは、nearレイアウトを使用し、2つのレプリカを持つアレイが作成されます。これら2つの値を変更するには、--layoutを使用します。[9.2.2.4項「YaSTおよびmdadmによるレプリカ数とレイアウトの指定」](#)を参照してください。

4. RAID 10デバイス/dev/md3上でファイルシステム(XFSファイルシステムなど)を作成します。

```
> sudo mkfs.xfs /dev/md3
```

これとは別のファイルシステムを使用するには、コマンドを変更します。

5. /etc/mdadm.confファイルを編集するか、ファイルがまだ存在しない場合は作成します(たとえば、[sudo vi /etc/mdadm.conf](#)を実行します)。次の行を追加します(ここでも、ファイルが既に存在する場合、最初の行は記述済みの可能性があります)。

```
DEVICE containers partitions
ARRAY /dev/md3 UUID=UUID
```

デバイスのUUIDは次のコマンドで取得できます。

```
> sudo mdadm -D /dev/md3 | grep UUID
```

6. /etc/fstabファイルを編集して、RAID 10デバイス/dev/md3のエントリを追加します。次の例は、XFSファイルシステム、およびマウントポイントとして/dataを使用するRAIDデバイスのエントリを示しています。

```
/dev/md3 /data xfs defaults 1 2
```

7. RAIDデバイスをマウントします。

```
> sudo mount /data
```

10 ディグレードRAIDアレイの作成

ディグレードアレイは、一部のデバイスが欠けたアレイです。ディグレードアレイは、RAID 1、RAID 4、RAID 5、およびRAID 6に対してのみサポートされています。これらのRAIDタイプは、その耐障害性機能として、一部のデバイスの欠落に耐えるように設計されています。通常、デバイスに障害が発生すると、ディグレードアレイが生成されます。ディグレードアレイは、意図的に作成するすることもできます。

RAIDの種類	許容可能な欠落スロット数
RAID 1	1つ以外の全スロット
RAID 4	1スロット
RAID 5	1スロット
RAID 6	1個または2個のスロット

一部のデバイスが欠落したディグレードアレイを作成するには、単に、デバイス名の代わりにmissingというワードを指定します。この指定により、mdadmは、アレイ内の対応するスロットを空のまま残します。

RAID 5アレイの作成時に、mdadmによって、余分なスペアドライブをもつディグレードアレイが自動的に作成されます。これは、一般に、ディグレードアレイ内にスペアを構築した方が、ディグレードアレイではないが正常でないアレイ上でパリティを再同期するより高速なためです。この機能は、--forceオプションで無効にできます。

RAIDを作成したいが、使用するデバイスの1つに既にデータが入っている場合は、ディグレードアレイを作成すると便利なことがあります。その場合は、他のデバイスでディグレードアレイを作成し、その使用中のデバイスからのデータをディグレードモードで実行中のRAIDにコピーし、デバイスをRAIDに追加して、RAIDの再構築まで待機すると、データがすべてのデバイスに行き渡ります。このプロセスの例を、次のプロシージャで示します。

1. 単一ドライブ//dev/sd1を使用してディグレードRAID 1デバイス//dev/md0を作成するには、コマンドプロンプトで、次のように入力します。

```
> sudo mdadm --create /dev/md0 -l 1 -n 2 /dev/sd1 missing
```

追加先のデバイスは、追加するデバイスと同じか、またはそれ以上のサイズをもつ必要があります。

- ミラーに追加したいデバイスに、RAIDアレイに移動したいデータが含まれている場合は、この時点で、そのデータを、ディグレードモードで実行中のRAIDアレイにコピーします。
- データのコピー元のデバイスをミラーに追加します。たとえば、/dev/sdb1をRAIDに追加するには、コマンドプロンプトで、次のように入力します。

```
> sudo mdadm /dev/md0 -a /dev/sdb1
```

一度に1つのデバイスのみ追加できます。カーネルがミラーを構築し、完全にオンラインにした後、別のミラーを追加できます。

- 構築の進捗状況を監視するには、コマンドプロンプトで、次のように入力します。

```
> sudo cat /proc/mdstat
```

毎秒更新されている間に再構築の進捗を確認するには、次のように入力します。

```
> sudo watch -n 1 cat /proc/mdstat
```

11 mdadmによるソフトウェアRAIDアレイのサイズ変更

本項では、ソフトウェアRAID 1、4、5、または6のデバイスのサイズを複数デバイス管理([mdadm\(8\)](#))ツールで増減する方法について説明します。

既存のソフトウェアRAIDデバイスのサイズ変更には、各コンポーネントパーティションが提供するスペースの増減が必要です。デバイスの使用可能な領域の変更を利用するためには、RAIDに存在するファイルシステムもサイズ変更できる必要があります。SUSE Linux Enterprise Serverでは、ファイルシステムBtrfs、Ext2、Ext3、Ext4、およびXFS(サイズの増加のみ)用のファイルシステムサイズ変更ユーティリティが提供されています。詳細については、[第2章 「ファイルシステムのサイズ変更」](#)を参照してください。

[mdadm](#)ツールは、ソフトウェアRAIDレベル1、4、5、および6に対してだけサイズ変更をサポートします。これらのRAIDレベルには耐障害性があるので、一度に1つずつ、サイズ変更するコンポーネントパーティションを削除できます。原則として、RAIDパーティションのホットリサイズが可能ですが、その場合は、データの保全に特に注意する必要があります。



警告: サイズ変更前のデータのバックアップ

パーティションまたはファイルシステムのサイズ変更には、データを失う可能性をはらむリスクが伴います。データの喪失を避けるには、データを必ずバックアップしてから、サイズ変更タスクを開始します。

RAIDのサイズ変更には、次のようなタスクがあります。タスクの実行順序は、サイズを増加するか、減少するかによって異なります。

表 11.1: RAIDのサイズ変更に必要なタスク

仕事	説明	サイズを 増大させ る場合の 順序	サイズを 減少させ る場合の 順序
各コンポーネントパーティションのサイズを変更します。	各コンポーネントパーティションのアクティブなサイズを増加または減少します。一度に1つのコンポーネントパーティションだけを削除し、そのサイズを変更してから、パーティションをRAIDに戻します。	1	2

仕事	説明	サイズを 増大させ る場合の 順序	サイズを 減少させ る場合の 順序
ソフトウェアRAID自体をサイズ変更します。	RAIDは、ベースのコンポーネントパーティションの増減を自動的には認識しません。したがって、RAIDに新しいサイズを知らせる必要があります。	2	3
ファイルシステムのサイズを変更します。	RAIDに常駐するファイルシステムをサイズ変更する必要があります。これは、サイズ変更のツールを提供するファイルシステムの場合のみ可能です。	3	1

以降の各項の手順では、次の表に示すデバイス名を使用します。これらの名前は変更して、必ずご使用のデバイスの名前を使用してください。

表 11.2: コンポーネントパーティションのサイズを増加するシナリオ

RAIDデバイス	コンポーネントパーティション
<u>/dev/md0</u>	<u>/dev/sda1</u>
	<u>/dev/sdb1</u>
	<u>/dev/sdc1</u>

11.1 ソフトウェアRAIDのサイズの増加

ソフトウェアRAIDのサイズを増やすには、複数のタスクを所定の順序で実行する必要があります。まずRAIDを構成するすべてのパーティションのサイズを増加させ、次にRAID自体のサイズを増加させます。そして最後に、ファイルシステムのサイズを増加させます。



警告: データ消失の可能性

RAIDに、ディスクの耐障害性がないか、単に一貫性がない場合、パーティションのどれかを削除すると、データが失われます。パーティションの削除は注意深く行い、必ず、データのバックアップをとってください。

11.1.1 コンポーネントパーティションのサイズの増加

RAID 1、4、5、または6のサイズを増加するには、本項の手順を適用します。RAID内のコンポーネントパーティションごとに、RAIDからパーティションを削除し、そのサイズを変更し、パーティションをRAIDに戻し、RAIDが安定するまで待機してから続行します。パーティションが削除されている間、RAIDはディグレードモードで動作し、ディスクの耐障害性がまったくないか、または低下しています。複数の同時ディスク障害を許容できるRAIDの場合でも、一度に2つ以上のパーティションを削除しないでください。RAID用コンポーネントパーティションのサイズを増加させるには、次の手順に従います。

1. 端末を開きます。
2. 次のように入力して、RAIDアレイが一貫性を保っており、同期されていることを確認します。

```
> cat /proc/mdstat
```

このコマンドの出力によって、RAIDアレイがまだ同期中とわかる場合は、同期化の完了まで待って、続行してください。

3. コンポーネントパーティションの1つをRAIDアレイから削除します。たとえば、次のように入力して、/dev/sda1を削除します。

```
> sudo mdadm /dev/md0 --fail /dev/sda1 --remove /dev/sda1
```

成功するためには、failとremoveの両方のアクションを指定する必要があります。

4. 次のオプションの1つを実行して、前の手順で削除したパーティションのサイズを増加させます。

- YaSTパーティショナやpartedなどのディスクパーティショナを使用して、パーティションのサイズを増やします。通常は、このオプションが選択されます。
- パーティションの常駐ディスクを、容量のより大きいデバイスに置き換えます。このオプションは、元ディスクの他のファイルシステムがシステムによりアクセスされない場合だけ選択できます。置き換え用デバイスをRAIDに追加すると、元のデバイスにあったデータをすべて再構築しなければならないので、データの同期にはるかに長い時間がかかります。

5. パーティションをRAIDアレイに再追加します。たとえば、次のように入力して、/dev/sda1を追加します。

```
> sudo mdadm -a /dev/md0 /dev/sda1
```

RAIDが同期され、一貫性をもつまで待機してから、次のパーティションの処理に進みます。

6. アレイ内の残りのコンポーネントデバイスごとに、これらの手順を繰り返します。正しいコンポーネントパーティションに対して、必ずコマンドを変更してください。
7. カーネルがRAIDのパーティションテーブルを再読み込みできないというメッセージが表示されたら、すべてのパーティションのサイズ変更後にコンピュータを再起動して、パーティションテーブルの更新を強制する必要があります。
8. [11.1.2項 「RAIDアレイのサイズの増加」](#)に進んでください。

11.1.2 RAIDアレイのサイズの増加

RAID内の各コンポーネントパーティションのサイズ変更後([11.1.1項 「コンポーネントパーティションのサイズの増加」](#)参照)も、新しい使用可能スペースの認識を強制するまで、RAIDアレイの設定では、元のアレイサイズが使用され続けます。RAIDアレイのサイズを指定したり、使用可能な最大スペースを使用できます。

本項の手順では、RAIDデバイスのデバイス名として`/dev/md0`を使用しています。この名前は変更して、必ずご使用のデバイスの名前を使用してください。

1. 端末を開きます。
2. 次のように入力して、RAIDアレイが一貫性を保っており、同期されていることを確認します。

```
> cat /proc/mdstat
```

このコマンドの出力によって、RAIDアレイがまだ同期中とわかる場合は、同期化の完了まで待って、続行してください。

3. 次のように入力して、アレイのサイズとアレイに認識されるデバイスサイズをチェックします。

```
> sudo mdadm -D /dev/md0 | grep -e "Array Size" -e "Dev Size"
```

4. 次のいずれかの操作を行います。

- 次のように入力して、アレイサイズを使用可能な最大サイズまで増加します。

```
> sudo mdadm --grow /dev/md0 -z max
```

- 次のように入力して、アレイサイズを使用可能な最大サイズまで増加します。

```
> sudo mdadm --grow /dev/md0 -z max --assume-clean
```

アレイは、デバイスに追加された領域を使用しますが、この領域は同期されません。これがRAID 1に推奨される理由は、同期が不要だからです。メンバー デバイスに追加されたスペースが事前にゼロ化されていれば、他のRAIDレベルに有用なことがあります。

- 次のように入力して、アレイサイズを指定の値まで増加します。

```
> sudo mdadm --grow /dev/md0 -z SIZE
```

SIZEを、キロバイト(1キロバイトは1024バイト)単位で目的のサイズを表す整数値で置き換えます。

5. 次のように入力して、アレイのサイズとアレイに認識されるデバイスサイズを再チェックします。

```
> sudo mdadm -D /dev/md0 | grep -e "Array Size" -e "Dev Size"
```

6. 次のいずれかの操作を行います。

- アレイのサイズ変更が成功していたら、[11.1.3項 「ファイルシステムのサイズの増加」](#)を続行します。
- アレイが予期どおりにサイズ変更されていない場合は、いったん再起動してから、このプロシージャを再試行する必要があります。

11.1.3 ファイルシステムのサイズの増加

アレイサイズの増加後は([11.1.2項 「RAIDアレイのサイズの増加」](#)参照)、ファイルシステムのサイズ変更ができます。

ファイルシステムのサイズを使用可能な最大スペースまで増加したり、正確なサイズを指定できます。ファイルシステムに正確なサイズを指定する場合は、その新しいサイズが次の条件を満たすかどうかを必ず確認してください。

- 新しいサイズは、既存データのサイズより大きくなければなりません。さもないと、データが失われます。
- ファイルシステムのサイズは使用可能なスペースより大きくできないので、新しいサイズは、現在のRAIDサイズ以下でなければなりません。

詳しい手順については、[第2章 「ファイルシステムのサイズ変更」](#)を参照してください。

11.2 ソフトウェアRAIDのサイズの削減

ソフトウェアRAIDのサイズを減らすには、複数のタスクを所定の順序で実行する必要があります。まずファイルシステムのサイズを縮小し、次にRAIDを構成するすべてのパーティションのサイズを縮小します。そして最後に、RAID自体のサイズを縮小します。



警告: データ消失の可能性

RAIDに、ディスクの耐障害性がないか、単に一貫性がない場合、パーティションのどちらかを削除すると、データが失われます。パーティションの削除は注意深く行い、必ず、データのバックアップをとってください。



重要: XFS

XFSでフォーマットされたファイルシステムのサイズを縮小することはできません。XFSではそのような機能がサポートされていないためです。そのため、XFSファイルシステムを使用するRAIDのサイズを縮小することはできません。

11.2.1 ファイルシステムのサイズの削減

RAIDデバイス上のファイルシステムのサイズを削減する際には、新しいサイズが次の条件を満たすかどうかを必ず確認してください。

- 新しいサイズは、既存データのサイズより大きくなければなりません。さもないと、データが失われます。
- ファイルシステムのサイズは使用可能なスペースより大きくできないので、新しいサイズは、現在のRAIDサイズ以下でなければなりません。

詳しい手順については、[第2章 「ファイルシステムのサイズ変更」](#)を参照してください。

11.2.2 RAIDアレイのサイズの削減

ファイルシステムのサイズ変更後([11.2.1項 「ファイルシステムのサイズの削減」](#)を参照)、RAIDアレイ設定では、利用可能スペースを縮小するよう強制するまで、元のアレイサイズを使い続けます。RAIDが、削減したセグメントサイズを使用するようにするには、mdadm --growモードを使用します。それを行うには、-zオプションを使用して、RAID内の各デバイ

スが使用するスペースの量を、キロバイトで指定する必要があります。このサイズは、チャンクサイズの倍数である必要があります。RAIDのスーパーブロックをデバイスに書き込むためのスペースとして、約128KBを残しておかなければなりません。

本項の手順では、RAIDデバイスのデバイス名として`/dev/md0`を使用しています。コマンドを変更して、必ずご使用のデバイスの名前を使用してください。

1. 端末を開きます。
2. 次のように入力して、アレイのサイズとアレイに認識されるデバイスサイズをチェックします。

```
> sudo mdadm -D /dev/md0 | grep -e "Array Size" -e "Dev Size"
```

3. 次のコマンドで、アレイのデバイスサイズを指定の値まで減少させます。

```
> sudo mdadm --grow /dev/md0 -z SIZE
```

`SIZE`を、キロバイト単位で目的のサイズを表す整数値で置き換えます。(1キロバイトは1024バイト)。

たとえば、次のコマンドでは、各RAIDデバイスのセグメントサイズを約40 GBに設定し、チャンクサイズは64 KBです。これには、RAIDのスーパーブロック用の128 KBが含まれます。

```
> sudo mdadm --grow /dev/md2 -z 41943168
```

4. 次のように入力して、アレイのサイズとアレイに認識されるデバイスサイズを再チェックします。

```
> sudo mdadm -D /dev/md0 | grep -e "Array Size" -e "Device Size"
```

5. 次のいずれかの操作を行います。

- アレイのサイズ変更が成功していたら、11.2.3項「コンポーネントパーティションのサイズの削減」を続行します。
- アレイが予期どおりにサイズ変更されていない場合は、いったん再起動してから、このプロシージャを再試行する必要があります。

11.2.3 コンポーネントパーティションのサイズの削減

RAID内の各デバイスで使用されるセグメントサイズの縮小後([11.2.2項 「RAIDアレイのサイズの削減](#) を参照)、各コンポーネントパーティション内の残りのスペースは、そのRAIDでは使われません。パーティションを現在のサイズのまま残して将来のRAIDの拡大に備えることも、今は使用しないそのスペースを利用するすることもできます。

そのスペースを利用するには、コンポーネントパーティションを1つずつ削減します。コンポーネントパーティションごとに、そのパーティションをRAIDから削除し、パーティションサイズを縮小し、パーティションをRAIDに戻したら、RAIDが安定するまで待機します。メタデータに備えるには、[11.2.2項 「RAIDアレイのサイズの削減](#) でRAIDに対して指定したサイズより、若干大きなサイズを指定する必要があります。

パーティションが削除されている間、RAIDはディグレードモードで動作し、ディスクの耐障害性がまったくないか、または低下しています。複数の同時ディスクエラーに耐えるRAIDの場合でも、一度に2つ以上のコンポーネントパーティションを削除しないでください。RAID用コンポーネントパーティションのサイズを縮小するには、次の手順に従います。

1. 端末を開きます。
2. 次のように入力して、RAIDアレイが一貫性を保っており、同期されていることを確認します。

```
> cat /proc/mdstat
```

このコマンドの出力によって、RAIDアレイがまだ同期中とわかる場合は、同期化の完了まで待って、続行してください。

3. コンポーネントパーティションの1つをRAIDアレイから削除します。たとえば、次のように入力して、/dev/sda1を削除します。

```
> sudo mdadm /dev/md0 --fail /dev/sda1 --remove /dev/sda1
```

成功するためには、failとremoveの両方のアクションを指定する必要があります。

4. 前の手順で削除したパーティションのサイズを、セグメントサイズに設定したサイズより若干小さいサイズに減らします。このサイズは、チャンクサイズの倍数であり、RAIDのスーパー ブロック用に128 KBを確保する必要があります。YaSTパーティショナやコマンドラインツールpartedなどを使用して、パーティションのサイズを縮小します。
5. パーティションをRAIDアレイに再追加します。たとえば、次のように入力して、/dev/sda1を追加します。

```
> sudo mdadm -a /dev/md0 /dev/sda1
```

RAIDが同期され、一貫性をもつまで待機してから、次のパーティションの処理に進みます。

6. アレイ内の残りのコンポーネントデバイスごとに、これらの手順を繰り返します。正しいコンポーネントパーティションに対して、必ずコマンドを変更してください。
7. カーネルがRAIDのパーティションテーブルを再読み込みできないというメッセージが表示されたら、すべてのパーティションのサイズ変更後にコンピュータを再起動する必要があります。
8. (オプション)RAIDとファイルシステムのサイズを拡大して、現在は小さめのコンポーネントパーティション内のスペースの最大量を利用し、後でファイルシステムのサイズを増やします。手順については、[11.1.2項 「RAIDアレイのサイズの増加」](#) を参照してください。

12 MDソフトウェアRAID用のストレージエンクロージャLEDユーティリティ

ストレージエンクロージャLEDモニタリングユーティリティ(`ledmon`)およびLEDコントロール(`ledctl`)ユーティリティは、多様なインターフェースおよびプロトコルを使用してストレージエンクロージャLEDを制御する、Linuxのユーザスペースアプリケーションです。その主たる用途は、`mdadm`ユーティリティで作成されたLinux MDソフトウェアのRAIDデバイスの状態を視覚化することです。`ledmon`デーモンがドライブアレイの状態を監視し、ドライブLEDの状態を更新します。`ledctl`ユーティリティを使用して、指定したデバイスに対するLEDパターンを設定できます。これらのLEDユーティリティでは、SGPIO (Serial General Purpose Input/Output)仕様(Small Form Factor (SFF) 8485)およびSCSI Enclosure Services (SES) 2プロトコルを使用して、LEDを制御します。SGPIO用のSFF-8489仕様のInternational Blinking Pattern Interpretation (IBPI)パターンを実装します。IBPIは、SGPIO規格がバックプレーン上のドライブやスロットの状態としてどのように解釈されるか、またバックプレーンがLEDでどのように状態を視覚化すべきかを定義します。

一部のストレージエンクロージャでは、SFF-8489仕様に厳格に準拠していないものがあります。エンクロージャプロセッサがIBPIパターンを受け入れていても、LEDの点滅はSFF-8489仕様に従っていない、あるいはプロセッサが限られた数のIBPIパターンしかサポートしていない場合があります。

LED管理(AHCI)およびSAF-TEプロトコルは、`ledmon`および`ledctl`ユーティリティではサポートされていません。

`ledmon`および`ledctl`アプリケーションは、インテルAHCIコントローラやインテルSASコントローラなどの、インテルのストレージコントローラで機能することが検証されています。MDソフトウェアのRAIDボリュームの一部であるPCIe-SSD(ソリッドステートドライブ)デバイスの、ストレージエンクロージャ状態(OK、Fail、Rebuilding)用LEDを制御するための、PCIe-SSD(ソリッドステートディスク)エンクロージャLEDもサポートされています。これらのアプリケーションは、他のベンダのIBPI準拠のストレージコントローラ(特にSAS/SCSIコントローラ)でも機能するはずですが、他のベンダのコントローラはテストされていません。

`ledmon`および`ledctl`は`ledmon`パッケージに付属しています。このパッケージはデフォルトではインストールされません。インストールするには、`sudo zypper in ledmon`を実行します。

12.1 ストレージエンクロージャLED監視サービス

ledmonアプリケーションは、MDソフトウェアRAIDデバイスの状態またはストレージエンクロージャまたはドライブベイ内のブロックデバイスの状態をコンスタントに監視する、デーモンプロセスです。一度に実行しているデーモンのインスタンスは、1つのみである必要があります。ledmonデーモンは、インテルのエンクロージャLEDユーティリティの一部です。

状態は、ストレージアレイエンクロージャまたはドライブベイ内の、各スロットに関連付けられたLED上で視覚化されます。このアプリケーションは、すべてのソフトウェアRAIDデバイスを監視し、その状態を視覚化します。選択したソフトウェアRAIDボリュームのみを監視する方法は、備わっていません。

ledmonデーモンでは、2種類のLEDシステム、すなわち、2 LEDシステム(Activity LEDとStatus LED)と、3 LEDシステム(Activity LED、Locate LED、およびFail LED)をサポートしています。このツールには、LEDへのアクセスの際に最高の優先度が与えられています。

ledmonを起動するには、次のように入力します。

```
> sudo ledmon [options]
```

[options]は次の1つ以上です。

ledmonのオプション

-c PATH,

--config=PATH

設定は`~/.ledctl`または`/etc/ledcfg.conf`(存在する場合)から読み込まれます。このオプションは、別の設定ファイルを指定する場合に使用します。

現時点では、複数の設定ファイルのサポートはまだ実装されていないため、このオプションは有効ではありません。詳細についてはman 5 ledctl.confを参照してください。

-l PATH,

--log=PATH

ローカルのログファイルへのパスを設定します。このユーザ定義ファイルを指定すると、グローバルログファイル`/var/log/ledmon.log`は使用されません。

-t SECONDS,

--interval=SECONDS

`sysfs`のスキャン間の時間間隔を設定します。値は秒単位です。最小値は5秒です。最大値の指定はありません。

--quiet, --error, --warning, --info, --debug, --all

詳細レベルを指定します。このレベルオプションは、情報なしから、ほとんどの情報までの順番で指定されます。ロギングを行わない場合は、--quietオプションを使用します。すべてをログする場合は、--allオプションを使用します。2つ以上の詳細オプションを指定した場合は、コマンド内の最後のオプションが適用されます。

-h,

--help

コマンド情報をコンソールに印刷して、終了します。

-v,

--version

ledmonのバージョンとライセンスに関する情報を表示して、終了します。



注記: 当バージョンの注意事項

ledmonデーモンは、SFF-8489仕様のPFA (Predicted Failure Analysis)状態は認識しません。したがって、PFAパターンは視覚化されません。

12.2 ストレージエンクロージャLED制御アプリケーション

エンクロージャLEDアプリケーション(ledctl)は、ストレージエンクロージャまたはドライブペイの各スロットに関連付けられたLEDを制御する、ユーザスペースアプリケーションです。ledctlアプリケーションは、インテルのエンクロージャLEDユーティリティの一部です。

このコマンドを発行すると、指定したデバイスのLEDが指定したパターンに設定され、それ以外のLEDはすべてオフになります。このアプリケーションはroot特権で実行する必要があります。ledmonアプリケーションはLEDへのアクセスに際して最高の優先度を持っているため、ledmonデーモンを実行中の場合は、ledctlで設定した一部のパターンが有効にならないことがあります(Locateパターン以外)。

ledctlアプリケーションでは、2種類のLEDシステム、すなわち、2LEDシステム(Activity LEDとStatus LED)と、3LEDシステム(Activity LED、Fail LED、およびLocate LED)をサポートしています。

ledctlを起動するには、次のように入力します。

```
> sudo [options] PATTERN_NAME=list_of_devices
```

[options]は次の1つ以上です。

-c PATH,

--config=PATH

ローカルの環境設定ファイルへのパスを設定します。このオプションを指定すると、グローバルの環境設定ファイルとユーザの環境設定ファイルは、無効になります。

-l PATH,

--log=PATH

ローカルのログファイルへのパスを設定します。このユーザ定義ファイルを指定すると、グローバルログファイル/var/log/ledmon.logは使用されません。

--quiet

stdoutまたはstderrに送信されるすべてのメッセージをオフにします。メッセージは、ローカルファイルおよびsyslogファシリティには引き続きログされます。

-h,

--help

コマンド情報をコンソールに印刷して、終了します。

-v,

--version

ledctlのバージョンとライセンスに関する情報を表示して、終了します。

12.2.1 パターン名

ledctlアプリケーションでは、SFF-8489仕様に従い、pattern_name引数に次の名前を使用できます。

locate

指定したデバイスまたはからのスロットに関連付けられたLocate LEDを点灯します。この状態は、スロットまたはドライブの識別に使用されます。

locate_off

指定したデバイスまたはからのスロットに関連付けられたLocate LEDを消灯します。

normal

指定したデバイスに関連付けられたStatus LED、Failure LED、およびLocate LEDを消灯します。

off

指定したデバイスに関連付けられたStatus LEDとFailure LEDのみを消灯します。

ica,

degraded

In a Critical Arrayパターンを視覚化します。

rebuild,

rebuild_p

Rebuildパターンを視覚化します。互換性とレガシーの理由から、両方のrebuild状態をサポートしています。

ifa,

failed_array

In a Failed Arrayパターンを視覚化します。

hotspare

Hotspareパターンを視覚化します。

pfa

Predicted Failure Analysisパターンを視覚化します。

failure,

disk_failed

Failureパターンを視覚化します。

ses_abort

SES-2 R/R ABORT

ses_rebuild

SES-2 REBUILD/REMAP

ses_ifa

SES-2 IN FAILED ARRAY

ses_ica

SES-2 IN CRITICAL ARRAY

ses_cons_check

SES-2 CONS CHECK

ses_hotspare

SES-2 HOTSPARE

ses_rsrd_dev

SES-2 RSVD DEVICE

ses_ok

SES-2 OK

ses_ident

SES-2 IDENT

ses_rm

SES-2 REMOVE

ses_insert

SES-2 INSERT

ses_missing

SES-2 MISSING

ses_dnr

SES-2 DO NOT REMOVE

ses_active

SES-2 ACTIVE

ses_enable_bb

SES-2 ENABLE BYP B

ses_enable_ba

SES-2 ENABLE BYP A

ses_devoff

SES-2 DEVICE OFF

ses_fault

SES-2 FAULT

非SES-2のパターンがエンクロージャ内のデバイスに送信されると、そのパターンは、上に示すように、SCSI Enclosure Services (SES) 2のパターンに自動的に変換されます。

表 12.1: 非SES-2パターンとSES-2パターン間での変換

非SES-2のパターン	SES-2のパターン
locate	ses_ident
locate_off	ses_ident
normal	ses_ok
off	ses_ok
ica	ses_ica
degraded	ses_ica
rebuild	ses_rebuild
rebuild_p	ses_rebuild
ifa	ses_ifa
failed_array	ses_ifa
hotspare	ses_hotspare
pfa	ses_rsvd_dev
failure	ses_fault
disk_failed	ses_fault

12.2.2 デバイスのリスト

ledctlコマンドを発行すると、指定したデバイスのLEDが指定したパターンに設定され、それ以外のLEDはすべてオフになります。デバイスのリストは、次の2つの形式のいずれかで提供できます。

- スペースなしのカンマで区切られたデバイスのリスト
- デバイスがスペースで区切られた波括弧内のリスト

同じコマンド内で複数のパターンを指定すると、各パターンに対するデバイスリストで、同一または異なるフォーマットを使用できます。2つのリスト形式を示す例は、[12.2.3項「例」](#) を参照してください。

デバイスは、`/dev`ディレクトリまたは`/sys/block`ディレクトリ内のファイルへのパスです。パスにより、ブロックデバイス、MDソフトウェアRAIDデバイス、またはコンテナデバイスを識別できます。ソフトウェアRAIDデバイスまたはコンテナデバイスの場合、報告されたLEDの状態は、関連付けられたブロックデバイスのすべてに対して設定されます。

`list_of_devices`にリストされているデバイスのLEDは、特定のパターンの`pattern_name`に設定され、それ以外のすべてのLEDは消灯されます。

12.2.3 例

単一のブロックデバイスを見つけるには

```
> sudo ledctl locate=/dev/sda
```

単一のブロックデバイスのLocate LEDを消灯するには

```
> sudo ledctl locate_off=/dev/sda
```

MDソフトウェアRAIDデバイスのディスクを見つけて、そのブロックデバイスの2つに同時に rebuildパターンを設定するには

```
> sudo ledctl locate=/dev/md127 rebuild={ /sys/block/sd[a-b] }
```

指定したデバイスに対するStatus LEDとFailure LEDを消灯するには

```
> sudo ledctl off={ /dev/sda /dev/sdb }
```

3つのブロックデバイスを見つけるには、次のいずれかのコマンドを実行します(どちらのコマンドでも同じです)。

```
> sudo ledctl locate=/dev/sda,/dev/sdb,/dev/sdc  
> sudo ledctl locate={ /dev/sda /dev/sdb /dev/sdc }
```

12.3 詳細情報

LEDのパターンおよび監視ツールに関する詳細は、次のリソースを参照してください。

- LEDMON open source project on GitHub.com (<https://github.com/intel/ledmon.git>) ↗

13 ソフトウェアRAIDのトラブルシューティング

/proc/mdstatファイルをチェックして、RAIDパーティションが破損しているかどうかを調べます。ディスクに障害が発生した場合、破損したハードディスクを、同じようにパーティション化されている新しいディスクに交換します。システムを再起動して、mdadm /dev/mdX --add /dev/sdXコマンドを入力します。Xを特定のデバイス識別子に置き換えてください。これにより、ハードディスクがRAIDシステムに自動的に統合され、そのRAIDシステムが完全に再構築されます(RAID 0を除くすべてのRAIDレベル)。

再構築中もすべてのデータにアクセスできますが、RAIDが完全に再構築されるまでは、パフォーマンスに問題が発生する場合があります。

13.1 ディスク障害復旧後の回復

RAIDアレイに含まれているディスクが障害を起こす理由はいくつかあります。最も一般的な理由を一覧にしました。

- ディスクメディアに問題が発生
- ディスクドライブコントローラに障害発生
- ディスクへの接続に障害発生

ディスクメディアまたはディスクコントローラの障害の場合、デバイスを交換または修理する必要があります。RAID内でホットスペアが設定されていない場合、手動による介入作業が必要です。

最後の接続障害の場合、接続の修復後(自動的に修復する場合もあります)、mdadmコマンドによって、障害が発生したデバイスは、自動的に再度追加されます。

md/mdadmは、ディスク障害の原因を正確に判断できないため、デバイスが正常であると明示的に指示されるまで、ディスクエラーを深刻なエラーと判断し、障害が発生しているデバイスを異常と見なします。

内部RAIDアレイを持つストレージデバイスなど、環境によっては、デバイス障害の原因の多くを接続の問題が占める場合があります。このような場合、mdadmに対して、デバイスが表示されたら、そのデバイスを--re-addによって自動的に再度追加しても問題ないと指示することができます。これには、以下の行を/etc/mdadm.confに追加します。

```
POLICY action=re-add
```

再表示されたらそのデバイスを自動的に再度追加できるのは、`udev`ルールによって、`mdadm -I DISK_DEVICE_NAME`が、自動的に表示されたあらゆるデバイスで実行されるように設定されている場合(デフォルトの動作)、および`write-intent`ビットマップが設定されている場合(デフォルトの設定)に限られることに注意してください。

このポリシーを特定のデバイスにのみ適用し、他には適用しない場合、`path=`オプションを/`etc/``mdadm.conf`内の`POLICY`行に追加して、選択したデバイスにのみデフォルトでないアクションを限定することができます。ワイルドカードを使用して、デバイスのグループを指定することができます。詳細については、`man 5 mdadm.conf`を参照してください。

IV ネットワークストレージ

- 14 Linux用iSNS 150
- 15 IPネットワークの大容量記憶域: iSCSI 158
- 16 Fibre Channel Storage over Ethernet Networks: FCoE 186
- 17 NVMe-oF 196
- 18 デバイスのマルチパスI/Oの管理 206
- 19 NFS共有ファイルシステム 260
- 20 Samba 282
- 21 autofsによるオンデマンドマウント 308

14 Linux用iSNS

ストレージエリアネットワーク(SAN)には、複数のネットワークにまたがる多数のディスクドライブを使用できます。これによって、デバイス検出とデバイスの所有権の判定が難しくなります。iSCSIイニシエータはSANのストレージリソースを識別し、どれにアクセスできるか判定できる必要があります。

Internet Storage Name Service(iSNS)は、TCP/IPネットワーク上のiSCSIデバイスの自動化された検出、管理、および設定を簡素化する、標準ベースのサービスです。iSNSでは、ファイバチャネルネットワークと同等の知的なストレージの検出および管理のサービスを提供します。

iSNSがない場合は、対象のターゲットが配置されている各ノードのホスト名またはIPアドレスを知っている必要があります。また、アクセス制御リストなどのメカニズムを使用して、どのイニシエータがどのターゲットにアクセスできるかを手動で管理する必要があります。

! 重要: セキュリティ上の考慮事項

ネットワークトラフィックが暗号化されていないため、iSNSは安全な内部ネットワーク環境でのみ使用される必要があります。

14.1 iSNSのしくみ

iSCSIイニシエータがiSCSIターゲットを検出するには、ネットワークのどのデバイスがストレージリソースで、アクセスするにはどのIPアドレスが必要かを特定する必要があります。iSNSサーバへクエリすると、iSCSIターゲットとイニシエータがアクセス許可をもつIPアドレスのリストが返されます。

iSNSを使用してiSNS検出ドメインを作成し、そこにiSCSIターゲットとiSCSIイニシエータをグループ化または構成します。多くのストレージノードを複数のドメインに振り分けることで、各ホストの検出プロセスをiSNSで登録された最適なターゲットのサブセットに限定でき、これによって、不要な検出を削減し、各ホストが検出関係の確立に費やす時間を制限することで、ストレージネットワークの規模を調整できるようになります。このようにして、ディスカバリ対象のターゲットとイニシエータの数を制御し、簡略化できます。

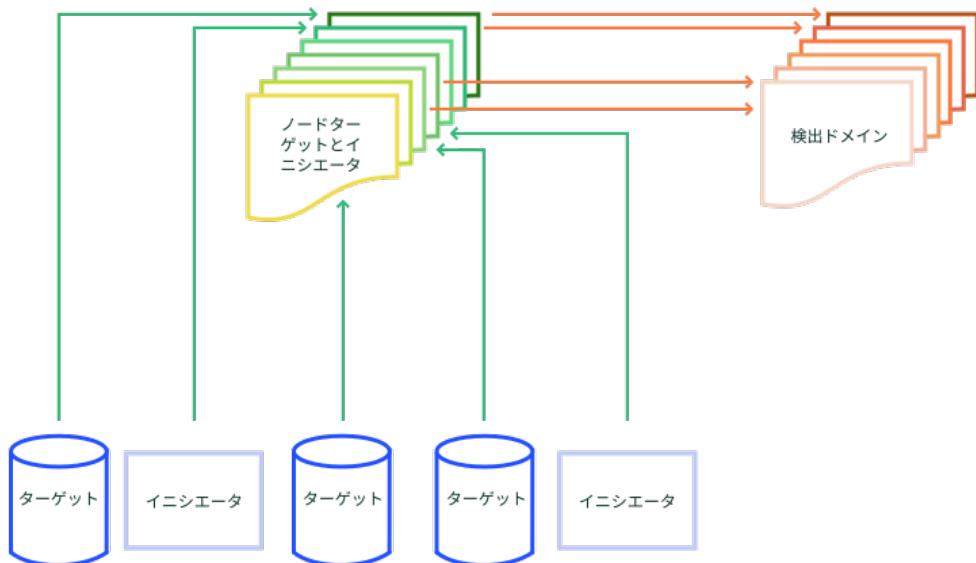


図 14.1: iSNS検出ドメイン

iSCSIターゲットとiSCSIイニシエータは両方とも、iSNSクライアントを使用して、iSNSプロトコルによるiSNSサーバとのトランザクションを開始できます。iSCSIターゲットとiSCSIイニシエータは、次にデバイス属性情報を共通検出ドメインに登録し、その他の登録されたクライアント情報をダウンロードし、検出ドメインで発生したイベントの非同期通知を受け取ります。

iSNSサーバは、iSNSプロトコルクエリとiSNSクライアントがiSNSプロトコルを使用して作成した要求に応答します。iSNSサーバはiSNSプロトコル状態変更通知を開始し、登録要求から送られてきた適切に認証された情報をiSNSデータベースに保存します。

Linux用iSNSは次の利点をもたらします。

- ネットワーク接続させたストレージ資産の登録、検出、管理に役立つ情報を提供する。
- DNSインフラストラクチャと統合する。
- iSCSIストレージの登録、検出、管理を統合する。
- ストレージ管理の実装が簡素化される。
- その他のディスカバリ方法よりもスケーラビリティが向上する。

iSNSにはいくつかの重要な利点があります。

たとえば、100個のiSCSIイニシエータと100個のiSCSIターゲットを使用したセットアップでは、すべてのiSCSIイニシエータが100個のiSCSIターゲットのいずれかを検出して接続しようとする可能性があります。イニシエータとターゲットをいくつかの検出ドメインにグループ化することで、ある部門のiSCSIイニシエータが別の部門のiSCSIターゲットを検出しないようにできます。

iSNSを使用する別の利点は、iSCSIクライアントが知っている必要があるのは、100台のサーバのホスト名またはIPアドレスではなく、iSNSサーバのホスト名またはIPアドレスだけであるということです。

14.2 Linux用iSNSサーバのインストール

Linux用iSNSサーバは、SUSE Linux Enterprise Serverに付属していますが、デフォルトではインストールも設定も行われません。パッケージ `open-isns` をインストールして、iSNSサービスを設定する必要があります。



注記: 同一サーバ上のiSNSとiSCSI

iSNSは、iSCSIターゲットまたはiSCSIイニシエータのソフトウェアがインストールされる同じサーバにインストールできます。ただし、iSCSIターゲットソフトウェアとiSCSIイニシエータソフトウェアの両方を同じサーバにインストールすることはできません。

Linux向けiSNSをインストールするには、次の手順に従います。

1. YaSTを起動してネットワークサービス › iSNSサーバを選択します。
2. `open-isns` がまだインストールされていない場合、今すぐインストールするようプロンプトが表示されます。インストールをクリックして確認します。
3. iSNSサービスの設定ダイアログが表示され、自動的にサービスタブが開きます。



4. サービスの開始で、次のいずれかを選択します。

- **起動時:** iSNSサービスは、サーバの起動時に自動的に開始します。
- **手動(デフォルト):** iSNSのインストール先サーバのサーバコンソールで、「sudo systemctl start isnsd」と入力して、iSNSサービスを手動で開始する必要があります。

5. 次のファイアウォール設定を指定します。

- **ファイアウォールでポートを開く:** このチェックボックスを選択して、ファイアウォールを開き、リモートコンピュータからサービスにアクセスできるようにします。ファイアウォールのポートは、デフォルトでは閉じています。
- **ファイアウォールの詳細:** ファイアウォールのポートを開いた場合、デフォルトでは、ポートがすべてのネットワークインターフェースで開きます。ポートを開くインターフェースを選択するには、Firewall Details(ファイアウォールの詳細)をクリックし、使用するネットワークインターフェースを選択し、次に、OKをクリックします。

6. OKをクリックして、設定を適用し、インストールを完了します。

7. 14.3項 「iSNS検出ドメインの設定」 に進んでください。

14.3 iSNS検出ドメインの設定

iSCSIイニシエータおよびターゲットでiSNSサービスを使用するには、これらが検出ドメインに属している必要があります。

! 重要: iSNSサービスがアクティブである必要がある

iSNS検出ドメインを設定するには、iSNSサービスがインストール済みで、実行されている必要があります。詳細については、[14.4項 「iSNSサービスの開始」](#) を参照してください。

14.3.1 iSNS検出ドメインの作成

iSNSサービスをインストールすると、デフォルトDDというデフォルトの検出ドメインが自動的に作成されます。iSNSを使用するように設定されている既存のiSCSIターゲットとイニシエータは、デフォルト検出ドメインに自動的に追加されます。

新しい検出ドメインを作成するには、次の手順に従います。

1. YaSTを起動して、ネットワークサービスの下でiSNSサーバを選択します。

2. 検出ドメインタブをクリックします。

検出ドメイン領域に既存のすべての検出ドメインが一覧にされます。Create Discovery Domains (検出ドメインの作成)で検出ドメインを作成したり、削除で既存の検出ドメインを削除したりできます。ドメインメンバーシップからiSCSIノードを削除すると、そのノードはドメインから削除されますが、iSCSIノード自体は削除されないことに注意してください。

検出ドメインメンバーの領域に、選択した検出ドメインに割り当てられているすべてのiSCSIノードがリストされます。別の検出ドメインを選択すると、その検出ドメインからのメンバーで、リストが更新されます。選択した検出ドメインからiSCSIノードを追加したり、削除できます。iSCSIノードを削除すると、そのノードは、ドメインから削除されますが、iSCSIノード自体は削除されません。

iSCSIノードメンバーの作成を使用すると、未登録のノードを検出ドメインのメンバーとして追加できます。iSCSIイニシエータまたはiSCSIターゲットがこのノードを登録すると、このノードは、このドメインの一部となります。

iSCSIイニシエータが検出要求を発行すると、iSNSサービスは同じ検出ドメイン内のメンバーであるすべてのiSCSIノードターゲットを返します。



3. 検出ドメインの作成ボタンをクリックします。
既存の検出ドメインを選択して削除ボタンをクリックして、その検出ドメインを削除できます。
4. 作成している検出ドメインの名前を指定して、OKをクリックします。
5. [14.3.2項 「iSCSIノードの検出ドメインへの追加」](#)に進んでください。

14.3.2 iSCSIノードの検出ドメインへの追加

1. YaSTを起動して、ネットワークサービスの下でiSNSサーバを選択します。
2. iSCSIノードタブをクリックします。



3. ノードのリストをレビューして、iSNSサービスを使用させたいiSCSIターゲットおよびイニシエータがリストされていることを確認します。
iSCSIターゲットまたはイニシエータが一覧にない場合、ノード上のiSCSIサービスを再起動する必要があります。それには以下を実行して、

```
> sudo systemctl restart iscsid.socket  
> sudo systemctl restart iscsi
```

イニシエータまたは

```
> sudo systemctl restart target-isns
```

ターゲットを再起動します。

iSCSIノードを選択して削除ボタンをクリックして、そのノードをiSNSデータベースから削除できます。iSCSIノードをもう使用しない場合や名前を変更した場合に有効です。iSCSI環境設定ファイルのiSNSの部分を削除したりコメント化していない限り、iSCSIノードは、iSCSIサービスの再開始時またはサーバの再起動時に、リスト(iSNSデータベース)に自動的に追加されます。

4. 検出ドメインタブをクリックして、目的の検出ドメインを選択します。
5. Add existing iSCSI Nodeをクリックしてドメインに追加するノードを選択し、ノードの追加をクリックします。

6. 検出ドメインに追加するノードの数だけ最後の手順を繰り返し、ノードの追加が終了したら完了をクリックします。
iSCSIノードは複数の検出ドメインに属することができます。

14.4 iSNSサービスの開始

iSNSは、インストール先のサーバで起動する必要があります。まだ起動時に開始するように設定していない場合(詳細については[14.2項 「Linux用iSNSサーバのインストール」](#)を参照)、端末で次のコマンドを入力します。

```
> sudo systemctl start isnsd
```

iSNSでは、stop、status、restartオプションも使用できます。

14.5 詳細情報

次のプロジェクトは、iSNSおよびiSCSIに関する詳細情報を提供します。

- iSNS server and client for Linux project (<https://github.com/open-iscsi/open-isns>) ↗
- iSNS client for the Linux LIO iSCSI target (<https://github.com/open-iscsi/target-isns>) ↗
- iSCSI tools for Linux (<https://www.open-iscsi.com>) ↗

iSNSの一般情報は、「RFC 4171: Internet Storage Name Service」 (<https://datatracker.ietf.org/doc/html/rfc4171>)に記載されています。

15 IPネットワークの大容量記憶域: iSCSI

コンピュータセンターや、サーバをサポートするサイトの主要タスクの1つは、適切なディスク容量を提供することです。この用途には、多くの場合、ファイバチャネルが使用されます。iSCSI(Internet SCSI)ソリューションは、ファイバチャネルに対する低コストの代替であり、コモディティサーバおよびEthernetネットワーキング装置を活用することができます。Linux iSCSIは、iSCSIイニシエータおよびiSCSI LIOターゲットのソフトウェアの提供により、Linuxサーバを中央ストレージシステムに接続します。

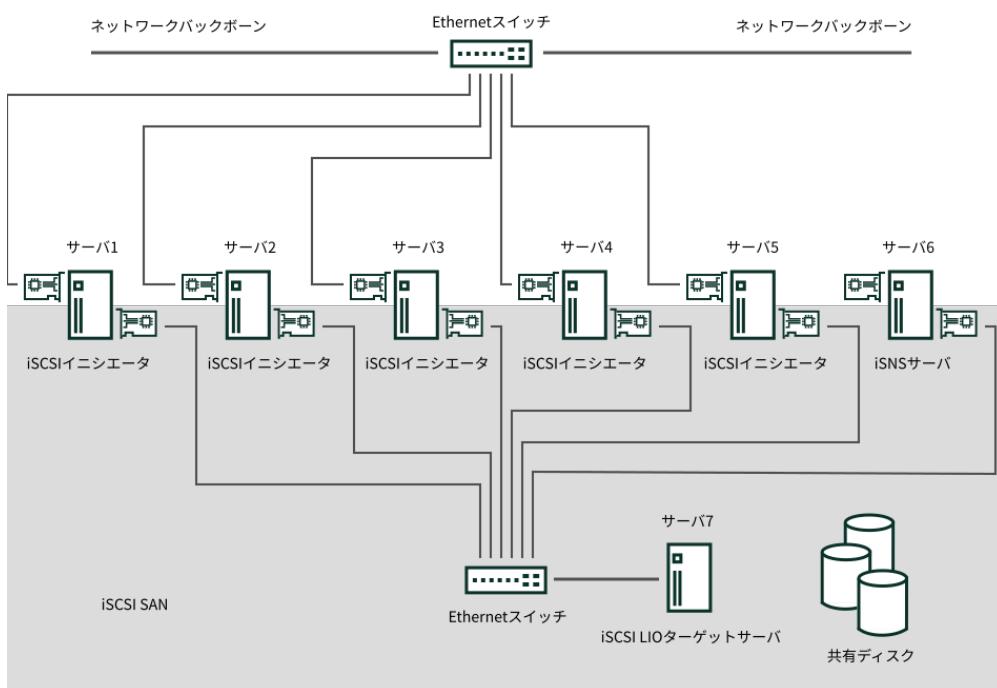


図 15.1: iSNSサーバによるiSCSI SAN



注記: LIO

LIOは、Linux用の標準のオープンソースマルチプロトコルSCSIターゲットです。LIOは、Linuxカーネルのバージョン2.6.38以降において、Linuxにおける標準の統一ストレージターゲットとして、STGT (SCSI Target)フレームワークにとって代わりました。SUSE Linux Enterprise Server 12では、古いバージョンのiSCSIターゲットサーバにiSCSI LIOターゲットサーバに代わっています。

iSCSIは、ストレージネットワーキングプロトコルであり、ブロックストレージデバイスとサーバ間における、TCP/IPネットワーク上でのSCSIパケットのデータ転送を簡素化にします。iSCSIターゲットソフトウェアは、ターゲットサーバ上で実行され、論理ユニットをiSCSI

ターゲットデバイスとして定義します。iSCSIイニシエータソフトウェアは異なるサーバ上で実行され、ターゲットデバイスに接続して、そのサーバ上でストレージデバイスを使用できるようにします。

iSCSI LIOターゲットサーバおよびiSCSIイニシエータサーバは、LAN内のIPレベルでSCSIパケットを送信して通信します。イニシエータサーバ上のアプリケーションがiSCSI LIOターゲットデバイスに対する照会を開始すると、オペレーティングシステムが必要なSCSIコマンドを発行します。するとSCSIコマンドが、**iSCSIイニシエータ**と呼ばれるソフトウェアによってIPパケットに組み込まれ、必要に応じて暗号化されます。パケットは内部IPネットワーク上で、「iSCSI LIOターゲットサーバ」または単に「iSCSIターゲット」と呼ばれる、対応するiSCSIリモートステーションに転送されます。

多くのストレージソリューションが、iSCSIによるアクセス手段を提供しています。また、LinuxサーバにiSCSIターゲットの役割をさせることもできます。この場合、Linuxサーバをファイルシステムサービス用に最適化しておくことが重要です。RAIDの詳細については、[第7章 「ソフトウェアRAIDの設定」](#)も参照してください。

15.1 iSCSI LIOターゲットサーバとiSCSIイニシエータのインストール

iSCSIイニシエータはデフォルトでインストールされますが(open-iscsiパッケージとyast2-iscsi-clientパッケージ)、iSCSI LIOターゲットパッケージは手動でインストールする必要があります。



重要: 同一サーバ上のイニシエータとターゲット

同一システムでイニシエータとターゲットを実行することはできますが、このセットアップは推奨されません。

iSCSI LIOターゲットサーバをインストールするには、端末で次のコマンドを実行します。

```
> sudo zypper in yast2-iscsi-lio-server
```

iSCSIイニシエータまたはその依存関係をインストールする必要がある場合は、コマンド **sudo zypper in yast2-iscsi-client** を実行します。

または、YaSTソフトウェア管理モジュールを使用してインストールします。

先に示したパッケージ以外の必要なパッケージは、インストーラによって自動的に組み込まれるか、またはそれぞれのYaSTモジュールの初回実行時にインストールされます。

15.2 iSCSI LIOターゲットサーバのセットアップ

本項では、YaSTを使用してiSCSI LIOターゲットサーバを構成し、iSCSI LIOのターゲットデバイスを設定する方法について説明します。任意のiSCSIイニシエータソフトウェアを使用して、ターゲットデバイスにアクセスすることができます。

15.2.1 iSCSI LIOターゲットサービスの起動およびファイアウォールの設定

iSCSI LIOターゲットサービスは、マニュアルで開始するようデフォルトで設定されています。同サービスを、システムのブート時に自動的に開始するよう設定できます。サーバでファイアウォールを使用していて、iSCSI LIOターゲットをほかのコンピュータでも利用可能としたい場合は、ターゲットへのアクセスに使用する各アダプタ用に、ファイアウォール内のポートを開放する必要があります。TCPポート3260が、iSCSIプロトコル用のポート番号です。これは、IANA (Internet Assigned Numbers Authority)により定義されています。

1. YaSTを起動し、ネットワークサービス › iSCSI LIOターゲットの順に起動します。
2. サービスタブに切り替えます。



3. サービスを開始で、iSCSI LIOターゲットサービスの開始方法を指定します。

- **起動時:** サービスは、サーバの再起動時に自動的に開始します。
- **手動:** (デフォルト)サーバの再起動後、`sudo systemctl start targetcli`コマンドを実行して、手動でサービスを開始する必要があります。ターゲットデバイスは、サービスを開始するまで利用できません。

4. サーバでファイアウォールを使用していて、iSCSI LIOターゲットをほかのコンピュータでも利用可能としたい場合は、ターゲットへのアクセスに使用する各アダプタインターフェース用に、ファイアウォール内のポート3260を開放します。このポートがネットワークインターフェースのすべてに対してクローズしている場合、iSCSI LIOターゲットはほかのコンピュータでは利用できません。

サーバでファイアウォールを使用していない場合、ファイアウォール設定は無効です。この場合、次の手順をスキップして、終了を使用して設定ダイアログから移動するか、別のタブに切り替えて設定を続行します。

- サービスタブで、ファイアウォールのポートを開くチェックボックスをオンにして、ファイアウォール設定を有効にします。
- ファイアウォールの詳細をクリックして、使用するネットワークインターフェースを確認または設定します。すべての利用可能なネットワークインターフェースが一覧表示され、デフォルトではすべてが選択されています。ポートを開く必要がないすべてのインターフェースを選択解除します。OKをクリックして設定を保存します。

5. 完了をクリックして、iSCSI LIOターゲットサービスの設定を保存して適します。

15.2.2 iSCSI LIOターゲットおよびイニシエータのディスカバリに対する認証の設定

iSCSI LIOターゲットサーバソフトウェアは、PPP-CHAP (Point-to-Point Protocol Challenge Handshake Authentication Protocol)をサポートしています。これは、Internet Engineering Task Force (IETF) RFC 1994 (<https://datatracker.ietf.org/doc/html/rfc1994>)で定義されている、3方向の認証方法です。サーバはこの認証方法を、ターゲット上のファイルへのアクセスにではなく、iSCSI LIOのターゲットとイニシエータのディスカバリ用に使用します。ディスカバリへのアクセスを制限しない場合は、認証なしを選択します。デフォルトでは検出認証なしオプションが有効になっています。このサーバ上のすべてのiSCSI LIOターゲットは、認証を要求しないので、同じネットワーク上のどのiSCSIイニシエータによっても検出することができます。

よりセキュアな設定に対する認証が必要な場合は、incoming認証、outgoing認証またはその両方を使用できます。イニシエータによる認証では、iSCSIイニシエータに、iSCSI LIOターゲット上で検出を実行するパーミッションがあることを証明するよう求めます。イニシエータは、incomingのユーザ名とパスワードを入力する必要があります。ターゲットによる認証では、iSCSI LIOターゲットに、自らが目的のターゲットであることをイニシエータに対して証明するよう求めます。iSCSI LIOターゲットは、outgoingのユーザ名とパスワードを、iSCSIイニシエータに提供する必要があります。パスワードはincomingとoutgoingのディスカバリで異なる必要があります。ディスカバリに対する認証を有効にしない場合、その設定は、すべてのiSCSI LIOターゲットグループに適用されます。

! 重要: セキュリティ

セキュリティ上の理由により、運用環境では、ターゲットおよびイニシエータのディスカバリに認証を使用することをお勧めします。

iSCSI LIOターゲットに対して認証の初期設定を行うには

1. YaSTを起動し、ネットワークサービス › iSCSI LIOターゲットの順に起動します。
2. グローバルタブに切り替えます。



3. デフォルトでは、認証は無効(検出認証なし)です。認証を有効にするには、イニシエータによる認証または送信認証、あるいはその両方を選択します。

4. 選択した認証方法に対して資格情報を提供します。ユーザ名とパスワードの組み合わせは、incomingとoutgoingディスカバリで異なっている必要があります。
5. 完了をクリックして、設定を保存して適用します。

15.2.3 ストレージスペースの準備

LUNをiSCSIターゲットサーバ用に設定する前に、使用するストレージを準備する必要があります。未フォーマットのブロックデバイス全体を1つのLUNとして使用することも、デバイスを複数の未フォーマットパーティションに再分割して、各パーティションを別個のLUNとして使用することもできます。iSCSIターゲット設定では、LUNをiSCSIイニシエータにエクスポートします。

YaSTのパーティショナまたはコマンドラインを使用して、パーティションを設定できます。詳細については、『展開ガイド』、第11章「エキスパートパーティショナ」、11.1項「熟練者向けパーティション設定の使用」を参照してください。iSCSI LIO ターゲットは、Linux、Linux LVM、またはLinux RAIDファイルシステムIDで未フォーマットのパーティションを使用できます。

! 重要: iSCSIターゲットデバイスをマウントしない

iSCSIターゲットとして使用するデバイスやパーティションを設定したら、ローカルパス経由で直接アクセスしないでください。ターゲットサーバにパーティションをマウントしないでください。

15.2.3.1 仮想環境でのデバイスのパーティション分割

仮想マシンのゲストサーバを、iSCSI LIOターゲットサーバとして使用できます。本項では、Xen仮想マシンにパーティションを割り当てる方法を説明します。また、SUSE Linux Enterprise Serverでサポートされている他の仮想環境も使用できます。

Xen仮想環境で、iSCSI LIOターゲットデバイスに使用するストレージスペースをゲストの仮想マシンに割り当て、ゲスト環境内の仮想ディスクとしてそのスペースにアクセスします。各仮想ディスクは、ディスク全体、パーティション、ボリュームなどの物理ブロックデバイスでも、Xenホストサーバ上の大規模な物理ディスク上の単一イメージファイルが仮想ディスクになっている、ファイルバックディスクイメージのいずれでも可能です。最適なパフォーマンスを得るために、物理ディスクまたはパーティションから各仮想ディスクを作成してください。ゲストの仮想マシンに仮想ディスクを設定したら、ゲストサーバを起動し、物理サーバの場合と同じ方法で、新しいブランクの仮想ディスクをiSCSIターゲットデバイスとして設定します。

ファイルバックディスクイメージがXenホストサーバ上に作成され、Xenゲストサーバに割り当てられます。デフォルトでは、Xenはファイルバックディスクイメージを`/var/lib/xen/images/VM_NAME`ディレクトリに保存します。ここで`VM_NAME`は仮想マシンの名前です。

15.2.4 iSCSI LIOターゲットグループの設定

YaSTを使用して、iSCSI LIOターゲットデバイスを設定することができます。YaSTは`targetcli`ソフトウェアを使用します。iSCSI LIOターゲットは、Linux、Linux LVM、またはLinux RAIDファイルシステムIDでパーティションを使用できます。

! 重要: パーティション

開始する前に、バックエンドストレージに使用するパーティションを選択します。パーティションをフォーマットする必要はありません。iSCSIクライアントは、接続時にそれらをフォーマットし、既存のすべてのフォーマットを上書きできます。

1. YaSTを起動し、ネットワークサービス › iSCSI LIOターゲットの順に起動します。
2. ターゲットタブに切り替えます。



3. 追加をクリックして、新しいiSCSI LIOのターゲットグループとデバイスを定義します。

iSCSI LIOターゲットソフトウェアにより、ターゲット、識別子、ポータルグループ、IPアドレス、およびポート番号の各フィールドが自動的に記入されます。認証を使用するが、デフォルトで選択されています。

- a. 複数のネットワークインターフェースがある場合は、[IPアドレス] ドロップダウンボックスを使用して、このターゲットグループ用に使用するネットワークインターフェースのIPアドレスを選択します。すべてのアドレスでサーバにアクセスできるようにするには、Bind All IP Addresses (すべてのIPアドレスをバインド)を選択します。
- b. このターゲットグループに対してイニシエータ認証を不要にする場合は、認証を使用をオフにします(非推奨)。
- c. Add (追加)をクリックします。デバイスまたはパーティションのパスを入力するか、または参照を使用して追加します。オプションで名前を指定して、OKをクリックします。0から始まるLUN番号が自動的に作成されます。フィールドを空にしておくと、名前が自動的に生成されます。
- d. (オプション)前の手順を繰り返し、このターゲットグループにターゲットを追加します。
- e. 目的のターゲットがすべてグループに追加されたら、次へをクリックします。

4. iSCSIターゲットイニシエータのセットアップの変更ページで、ターゲットグループ内のLUNへのアクセスを許可されるイニシエータに関する情報を設定します。

iSCSIターゲットイニシエータのセットアップの変更



ターゲットグループに対して少なくとも1つ以上のイニシエータを指定すると、LUNの編集、認証の編集、削除、およびコピーの各ボタンが有効になります。追加またはコピーを使用して、ターゲットグループにイニシエータを追加できます。

[iSCSIターゲットの変更] : オプション

- **追加:** 選択したiSCSI LIOターゲットグループに、新たなイニシエータのエントリを追加します。
- **LUNを編集:** iSCSI LIOターゲットグループ内のどのLUNが、選択したイニシエータにマップするかを設定します。割り当てられたターゲットのそれぞれを、任意のイニシエータにマップすることができます。
- **認証を編集:** 選択したイニシエータに対する好みの認証方法を設定します。認証なしを指定することも、incoming認証、outgoing認証、またはその両方を設定することもできます。
- **削除:** 選択したイニシエータのエントリを、ターゲットグループに割り当てられたイニシエータのリストから削除します。
- **コピー:** 同じLUNのマッピングと認証設定を持つ新たなイニシエータのエントリを、選択したイニシエータのエントリとして追加します。これにより、容易に同じ共有LUNを、クラスタ内のがノードに順々に割り当てることができます。

- a. 追加をクリックして、イニシエータ名を指定し、TPGからLUNをインポートチェックボックスをオンまたはオフにしてから、OKをクリックして設定を保存します。
- b. イニシエータのエントリを選択して、LUNの編集をクリックし、LUNのマッピングを変更してiSCSI LIOターゲットグループ内のどのLUNを選択したイニシエータに割り当てるかを指定して、OKをクリックして変更内容を保存します。
iSCSI LIOターゲットグループが複数のLUNで構成されている場合は、1つまたは複数のLUNを、選択したイニシエータに割り当てることができます。デフォルトでは、グループ内の使用可能なLUNのそれぞれが、イニシエータLUNに割り当てられます。

LUNの割り当てを変更するには、次の操作の1つ以上を実行します。

- **追加:** 追加をクリックして新しいイニシエータのLUNのエントリを作成し、変更ドロップダウンボックスを使用して、そのエントリにターゲットLUNをマップします。
- **削除:** イニシエータのLUNのエントリを選択し、削除をクリックしてターゲットLUNのマッピングを削除します。
- **変更:** イニシエータのLUNのエントリを選択し、変更ドロップダウンボックスを使用して、そのエントリにマップするターゲットLUNを選択します。

一般的な割り当つのプランには、次のようなものがあります。

- 1台のサーバが、イニシエータとして登録されています。ターゲットグループ内のLUNがすべて、それに割り当てられています。
このグループ化戦略を使用して、特定のサーバに対して、iSCSI SANストレージを論理的にグループ化することができます。
- 複数の独立したサーバが、イニシエータとして登録されています。1つまたは複数のターゲットLUNが、それぞれのサーバに割り当てられています。それぞれのLUNは、1台のサーバのみに割り当てられています。
このグループ化戦略を使用して、データセンター内の特定の部門またはサービスのカテゴリに対して、iSCSI SANストレージを論理的にグループ化することができます。
- クラスタの各ノードが、イニシエータとして登録されています。共有のターゲットLUNがすべて、各ノードに割り当てられています。すべてのノードがデバイスに接続されていますが、ほとんどのファイルシステムに対して、クラスタソフトウェアによってデバイスによるアクセスがロックされ、一度に1つのノード上にのみデバイスがマウントされます。共有ファイルシステム

(OCFS2など)では、複数のノードが同時に同じファイル構造をマウントし、読み込みおよび書き込みアクセスを持つ同じファイルを開くことが可能です。

このグループ化戦略を使用して、特定のサーバクラスタに対して、iSCSI SANストレージを論理的にグループ化することができます。

- c. イニシエータのエントリを選択して、認証の編集をクリックし、イニシエータに対する認証設定を指定してから、OKをクリックして設定を保存します。検出認証なしとすることも、イニシエータによる認証、送信認証、または両方を設定することもできます。各イニシエータに対して指定できるユーザ名とパスワードの組み合わせは、1つだけです。イニシエータに対するincoming認証とoutgoing認証の資格情報は、異なっても構いません。資格情報は、イニシエータごとに異なっても構いません。
- d. このターゲットグループにアクセスできる各iSCSIイニシエータについて、前の手順を繰り返します。
- e. イニシエータの割り当てを設定し終わったら、次へをクリックします。

5. 完了をクリックして、設定を保存して適用します。

15.2.5 iSCSI LIOターゲットグループの変更

以下のようにして、iSCSI LIOターゲットグループに変更を加えることができます。

- ターゲットLUNデバイスをターゲットグループに追加または削除する
- ターゲットグループに対してイニシエータを追加または削除する
- ターゲットグループのイニシエータに対する、イニシエータLUNからターゲットLUNへのマッピングを変更する
- イニシエータ認証(incoming、outgoing、または両方)用のユーザ名とパスワードの資格情報を変更する

iSCSI LIOターゲットグループに対する設定を確認または変更するには:

1. YaSTを起動し、ネットワークサービス › iSCSI LIOターゲットの順に起動します。
2. ターゲットタブに切り替えます。
3. 変更するiSCSI LIOターゲットグループを選択して、編集をクリックします。

4. [iSCSIターゲットLUNのセットアップを変更] ページで、ターゲットグループにLUNを追加し、LUNの割り当てを編集するか、またはターゲットLUNをグループから削除します。すべてグループに目的の変更が行われたら、次へをクリックします。
オプション情報については、[\[iSCSIターゲットの変更\]](#) : オプションを参照してください。
5. [iSCSIターゲットイニシエータのセットアップの変更] ページで、ターゲットグループ内のLUNへのアクセスを許可されるイニシエータに関する情報を設定します。すべてグループに目的の変更が行われたら、次へをクリックします。
6. 完了をクリックして、設定を保存して適用します。

15.2.6 iSCSI LIOターゲットグループの削除

iSCSI LIOターゲットグループを削除すると、グループの定義と、イニシエータに対する関連のセットアップ(LUNのマッピングや認証資格情報を含む)が削除されます。パーティション上のデータは破棄されません。イニシエータに再度アクセス権を付与するには、ターゲットLUNを別のターゲットグループまたは新規のターゲットグループに割り当てて、それらに対するイニシエータアクセスを設定します。

1. YaSTを起動し、ネットワークサービス → iSCSI LIOターゲットの順に起動します。
2. ターゲットタブに切り替えます。
3. 削除するiSCSI LIOターゲットグループを選択して、削除をクリックします。
4. 確認のメッセージが表示されたら、続行をクリックして削除を確認するか、キャンセルをクリックしてキャンセルします。
5. 完了をクリックして、設定を保存して適用します。

15.3 iSCSIイニシエータの設定

iSCSIイニシエータを使用して、任意のiSCSIターゲットに接続できます。これは、[15.2項「iSCSI LIOターゲットサーバのセットアップ」](#)で説明されているターゲットソリューションだけに限りません。iSCSIイニシエータの設定には、利用可能なiSCSIターゲットの検出と、iSCSIセッションの設定という2つの主要ステップがあります。どちらの設定も、YaSTを使って行うことができます。

15.3.1 YaSTを使ったiSCSIイニシエータの設定

YaSTの [iSCSIイニシエータの概要] が3つのタブに分割されます。

サービス:

サービスタブでは、ブート時にiSCSIイニシエータを有効にできます。固有のイニシエータ名とディスカバリに使用するiSNSサーバも設定できます。

接続したターゲット:

Connected Targetsタブには、現在接続しているiSCSIターゲットの概要が表示されます。このタブにも、検出されたターゲットタブのように、システムに新しいターゲットを追加するオプションが用意されています。

検出されたターゲット:

検出されたターゲットタブでは、ネットワーク内のiSCSIターゲットを手動で検出することができます。

15.3.1.1 iSCSIイニシエータの設定

1. YaSTを起動し、ネットワークサービス › iSCSIイニシエータ の順に起動します。
2. サービスタブに切り替えます。



3. 設定の書き込み後で、設定変更があった場合に何をするかを定義します。選択できるオプションはサービスの現在のステータスによって異なります。

現在の状態を維持オプションを選択すると、サービスは同じ状態のままでです。

4. 再起動後メニューでは、再起動後に実行するアクションを指定します。

- 起動時に開始 - 起動時にサービスを自動的に開始します。
- オンデマンドで開始 - 関連するソケットが実行され、必要に応じてサービスが開始されます。
- 開始しない - サービスは自動的には開始しません。
- 現在の設定を維持 - サービス設定は変更されません。

5. イニシエータ名を指定、または確認します。

このサーバ上のiSCSIイニシエータに、正しい形式のiSCSI修飾名(iQN)を指定します。イニシエータ名はネットワーク全体で固有のものでなければなりません。IQNは次の一般的なフォーマットを使用します。

```
iqn.yyyy-mm.com.mycompany:n1:n2
```

ここでn1とn2はアルファベットか数字です。例:

```
iqn.1996-04.de.suse:01:a5dfcea717a
```

イニシエータ名には、サーバ上の`/etc/iscsi/initiatorname.iscsi`ファイルから対応する値が自動的に入力されます。

サーバがiBFT(iSCSI Boot Firmware Table)をサポートしている場合は、イニシエータ名にはIBFT内の対応する値が入力され、このインターフェースではイニシエータ名を変更できません。代わりにBIOSセットアップを使用して変更してください。iBFTは、サーバのiSCSIターゲットとイニシエータの説明を含む、iSCSIの起動プロセスに便利な各種パラメータを含んだ情報ブロックです。

6. 次のいずれかの方法を使用して、ネットワーク上のiSCSIターゲットを検出します。

- **iSNS:** iSNS (Internet Storage Name Service)を使用してiSCSIターゲットを検出するには、続いて[15.3.1.2項 「iSNSによるiSCSIターゲットの検出」](#)を実行します。
- **検出されたターゲット:** iSCSIターゲットデバイスを手動で検出するには、続いて[15.3.1.3項 「iSCSIターゲットの手動検出」](#)を実行します。

15.3.1.2 iSNSによるiSCSIターゲットの検出

このオプションを使用する前に、ご使用の環境内でiSNSサーバをインストールし、設定しておく必要があります。詳細については、[第14章 「Linux用iSNS」](#)を参照してください。

1. YaSTでiSCSIイニシエータを選択し、次にサービスタブを選択します。
2. iNSNサーバのIPアドレスとポートを指定します。デフォルトポートは3205です。
3. OKをクリックして、変更内容を保存して適用します。

15.3.1.3 iSCSIターゲットの手動検出

iSCSIイニシエータを設定しているサーバからアクセスする各iSCSIターゲットサーバについて、次の手順を繰り返し実行します。

1. YaSTでiSCSIイニシエータを選択し、次に検出されたターゲットタブを選択します。
2. 検出をクリックして [iSCSIイニシエータの検出] ダイアログを開きます。
3. IPアドレスを入力し、必要に応じてポートを変更します。デフォルトポートは3260です。
4. 認証が必要な場合は、検出認証なしをオフにして、イニシエータによる認証またはターゲットによる認証で資格情報を指定します。
5. 次へをクリックして、検出を開始し、iSCSIターゲットサーバに接続します。
6. 資格情報が必要な場合は、検出成功後、接続を使用してターゲットを有効化します。指定したiSCSIターゲットを使用するための、認証資格情報の提供を促されます。
7. 次へをクリックして、設定を完了します
これでターゲットが接続したターゲットに表示され、仮想iSCSIデバイスが使用可能になります。
8. OKをクリックして、変更内容を保存して適用します。
9. `lsscsi`コマンドを使用すると、iSCSIターゲットデバイスのローカルデバイスパスを検出することができます。

15.3.1.4 iSCSIターゲットデバイスの起動設定

1. YaSTで、iSCSIイニシエータを選択し、次に接続したターゲットタブを選択して、現在サーバに接続されているiSCSIターゲットデバイスの一覧を表示することができます。
2. 管理するiSCSIターゲットデバイスを選択します。
3. 起動の切り替えをクリックして設定を変更します。

自動: このオプションは、iSCSIサービス自体の起動時に接続するiSCSIターゲットに使用されます。これが通常の設定です。

Onboot(起動時): このオプションは、起動時、つまりルート(/)がiSCSI上にある場合に接続するiSCSIターゲットに使用します。したがって、iSCSIターゲットデバイスはサーバの起動時にinitrdによって評価されます。このオプションはIBM Zなど、iSCSIからブートできないプラットフォームでは無視されます。したがって、これらのプラットフォームでは使用しないでください。代わりに自動を使用してください。

4. OKをクリックして、変更内容を保存して適用します。

15.3.2 手動によるiSCSIイニシエータの設定

iSCSI接続の検出や設定を行うには、iscsidが稼働していなければなりません。初めてディスカバリを実行する場合、iSCSIイニシエータの内部データベースが/etc/iscsi/ディレクトリに作成されます。

ディスカバリがパスワードにより保護されている場合は、iscsidに認証情報を渡します。最初にディスカバリを実行するときには内部データベースが存在していないため、現時点でこれは使用できません。代わりに、/etc/iscsid.conf設定ファイルを編集して、情報を指定する必要があります。ディスカバリのパスワード情報を追加するには、/etc/iscsid.confファイルの最後に、次の行を追加します。

```
discovery.sendtargets.auth.authmethod = CHAP
discovery.sendtargets.auth.username = USERNAME
discovery.sendtargets.auth.password = PASSWORD
```

ディスカバリは、受け取ったすべての値を内部データベースに保存します。また、検出したターゲットをすべて表示します。次のコマンドで、このディスカバリを実行します。

```
> sudo iscsiadm -m discovery --type=st --portal=TARGET_IP
```

次のように出力されます。

```
10.44.171.99:3260,1 iqn.2006-02.com.example.iserv:systems
```

iSNSサーバで使用できるターゲットを検出するには、次のコマンドを使用します。

```
sudo iscsiadm --mode discovery --type isns --portal TARGET_IP
```

iSCSIターゲットに定義されている各ターゲットが、それぞれ1行に表示されます。保存されたデータの詳細については、15.3.3項「iSCSIイニシエータデータベース」を参照してください。

`iscsiadm`コマンドの`--login`オプションを使用すると、必要なすべてのデバイスが作成されます。

```
> sudo iscsiadm -m node -n iqn.2006-02.com.example.iserv:systems --login
```

新しく生成されたデバイスは`lsscsi`コマンドの出力に表示され、マウントできるようになります。

15.3.3 iSCSIイニシエータデータベース

iSCSIイニシエータにより検出されたあらゆる情報は、`/etc/iscsi`に存在する2つのデータベースファイルに保存されます。1つは、ディスカバリが検出したターゲット用のデータベースで、もう1つは検出したノード用のデータベースです。データベースにアクセスする場合、まずデータをディスカバリ用データベースから取得するのか、またはノードデータベースから取得するのかを指定する必要があります。これには、`iscsiadm`の`-m discovery`および`-m node`パラメータを使用します。`iscsiadm`コマンドに、どちらかのパラメータを指定して実行すると、そのデータベースに保管されているレコードの概要が表示されます。

```
> sudo iscsiadm -m discovery  
10.44.171.99:3260,1 iqn.2006-02.com.example.iserv:systems
```

この例のターゲット名は`iqn.2006-02.com.example.iserv:systems`です。このデータセットに関する操作を行う場合に、この名前が必要になります。ID `iqn.2006-02.com.example.iserv:systems`のデータレコードのコンテンツを調べるには、次のコマンドを使用します。

```
> sudo iscsiadm -m node --targetname iqn.2006-02.com.example.iserv:systems  
node.name = iqn.2006-02.com.example.iserv:systems  
node.transport_name = tcp  
node.tpgt = 1  
node.active_conn = 1  
node.startup = manual  
node.session.initial_cmds = 0  
node.session.reopen_max = 32  
node.session.auth.authmethod = CHAP  
node.session.auth.username = joe  
node.session.auth.password = *****  
node.session.auth.username_in = EMPTY  
node.session.auth.password_in = EMPTY  
node.session.timeo.replacement_timeout = 0  
node.session.err_timeo.abort_timeout = 10  
node.session.err_timeo.reset_timeout = 30  
node.session.iscsi.InitialR2T = No  
node.session.iscsi.ImmediateData = Yes
```

....

これらの変数の値を変更する場合は、iscsiadmコマンドでupdateオプションを使用します。たとえば、初期化時にiscidをiSCSIターゲットにログインさせる場合は、値にautomaticとnode.startupを設定します。

```
sudo iscsiadm -m node -n iqn.2006-02.com.example.iserv:systems \
-p ip:port --op=update --name=node.startup --value=automatic
```

不要になったデータセットを削除する場合は、delete操作を使用します。ターゲット*iqn.2006-02.com.example.iserv:systems*が有効なレコードではなくなった場合は、このレコードを次のコマンドで削除します。

```
> sudo iscsiadm -m node -n iqn.2006-02.com.example.iserv:systems \
-p ip:port --op=delete
```

! 重要: 確認は表示されない

このオプションでは、確認のメッセージを表示せずにレコードを削除するため、使用する際には細心の注意を払うようしてください。

検出したすべてのターゲットのリストを取得するには、sudo iscsiadm -m nodeコマンドを実行します。

15.4 targetcli-fbを使用したソフトウェアターゲットの設定

targetcliは、LinuxIO (LIO) ターゲットサブシステムの設定を管理するためのシェルです。シェルは対話的に呼び出すことも、従来のシェルと同様に一度に1つのコマンドを実行することもできます。従来のシェルと同様に、cdコマンドを使用してtargetcli機能階層をトラバースし、コンテンツをlsコマンドで一覧表示します。

使用可能なコマンドは、現在のディレクトリによって異なります。各ディレクトリには独自のコマンドセットがありますが、すべてのディレクトリで使用可能なコマンドもあります(たとえば、cdとlsコマンド)。

targetcliコマンドは次のフォーマットを持ちます。

```
[DIRECTORY] command [ARGUMENTS]
```

任意のディレクトリでhelpコマンドを使用して、使用可能なコマンドのリスト、または特定のコマンドに関する情報を表示できます。

targetcliツールは、targetcli-fbパッケージの一部です。このパッケージは公式のSUSE Linux Enterprise Server ソフトウェアリポジトリで入手でき、次のコマンドを使用してインストールできます。

```
> sudo zypper install targetcli-fb
```

targetcli-fbパッケージがインストールされたら、targetcliサービスを有効にします。

```
> sudo systemctl enable targetcli  
> sudo systemctl start targetcli
```

targetcliシェルに切り替えるには、ルートとしてtargetcliを実行します。

```
> sudo targetcli
```

デフォルトの設定を確認するには、lsコマンドを実行できます。

```
/> ls  
o- / ..... [....]  
o- backstores ..... [....]  
| o- block ..... [Storage Objects: 0]  
| o- fileio ..... [Storage Objects: 0]  
| o- pscsi ..... [Storage Objects: 0]  
| o- ramdisk ... [Storage Objects: 0]  
| o- rbd ..... [Storage Objects: 0]  
o- iscsi ..... [Targets: 0]  
o- loopback ..... [Targets: 0]  
o- vhost ..... [Targets: 0]  
o- xen-pvscsi ..... [Targets: 0]  
/>
```

lsコマンドの出力はバックエンドが設定されていないことを示します。したがって、最初の手順はサポートされているソフトウェアターゲットの1つを設定することです。

targetcliは次のバックエンドをサポートしています。

- fileio、ローカルイメージファイル
- block、専用ディスクまたはパーティション上のブロックストレージ
- pscsi、SCSIバススルーデバイス
- ramdisk、メモリベースのバックエンド
- rbd、Ceph RADOSブロックデバイス

targetcliの機能を理解するには、createコマンドを使用してソフトウェアターゲットとしてローカルイメージファイルを設定します。

```
/backstores/fileio create test-disc /alt/test.img 1G
```

これにより、指定された場所(この場合は`/alt`)に1 GBの`test.img`イメージが作成されます。`ls`を実行すると、次の結果が表示されます。

```
/> ls
o- /
  o- backstores ..... [....]
    | o- block ..... [Storage Objects: 0]
    | o- fileio ..... [Storage Objects: 1]
    | | o- test-disc ... [/alt/test.img (1.0GiB) write-back deactivated]
    | |   o- alua ..... [ALUA Groups: 1]
    | |     o- default_tg_pt_gp .... [ALUA state: Active/optimized]
    | o- pscsi ..... [Storage Objects: 0]
    | o- ramdisk ..... [Storage Objects: 0]
    | o- rbd ..... [Storage Objects: 0]
  o- iscsi ..... [Targets: 0]
  o- loopback ..... [Targets: 0]
  o- vhost ..... [Targets: 0]
  o- xen-pvscsi ..... [Targets: 0]
/>
```

出力は、`/backstores/fileio`ディレクトリの下に、作成された`/alt/test.img`というファイルにリンクされている`test-disc`と呼ばれるファイルベースのバックストアがあることを示しています。新しいバックストアはまだ有効になっていないことに注意してください。

次の手順は、iSCSIターゲットのフロントエンドをバックエンドストレージに接続することです。各ターゲットには、IQN(iSCSI修飾名)が必要です。最も一般的に使用されるIQN形式は次のとおりです。

```
iqn.YYYY-MM.NAMING-AUTHORITY:UNIQUE-NAME
```

IQNの次の部分が必要です。

- YYYY-MM、命名機関が設立された年と月
- NAMING-AUTHORITY、命名機関のインターネットドメイン名の逆構文
- UNIQUE-NAME、命名機関によって選択されたドメイン固有の名前

たとえば、ドメイン`open-iscsi.com`の場合、IQNは次のようにになります。

```
iqn.2005-03.com.open-iscsi:UNIQUE-NAME
```

iSCSIターゲットの作成時に、targetcliコマンドを使用すると、指定された形式に従っている限り、独自のIQNを割り当てることができます。たとえば、次のように、ターゲットの作成時に名前を省略して、コマンドでIQNを作成することもできます。

```
/> iscsi/ create
```

`ls`コマンドを再び実行します。

```

/> ls
o- / ..... [....]
o- backstores ..... [....]
| o- block ..... [Storage Objects: 0]
| o- fileio ..... [Storage Objects: 1]
| | o- test-disc ..... [/alt/test.img (1.0GiB) write-back deactivated]
| | | o- alua ..... [ALUA Groups: 1]
| | | o- default_tg_pt_gp ..... [ALUA state: Active/optimized]
| o- pscsi ..... [Storage Objects: 0]
| o- ramdisk ..... [Storage Objects: 0]
| o- rbd ..... [Storage Objects: 0]
o- iscsi ..... [Targets: 1]
| o- iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456 ... [TPGs: 1]
| | o- tpg1 ..... [no-gen-acls, no-auth]
| | | o- acls ..... [ACLs: 0]
| | | o- luns ..... [LUNs: 0]
| | | o- portals ..... [Portals: 1]
| | | | o- 0.0.0.0:3260 ..... [OK]
o- loopback ..... [Targets: 0]
o- vhost ..... [Targets: 0]
o- xen-pvscsi ..... [Targets: 0]
/>

```

出力には、自動的に生成された iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456iqn.を持つ作成されたiSCSIターゲットノードが表示されます

targetcliでは、デフォルトのターゲットポータルグループtpg1も作成し、有効にしていることに注意してください。これは、ルートレベルの変数auto_add_default_portalとauto_enable_tpgtがデフォルトでtrueに設定されているために実行されます。

このコマンドは、0.0.0.0IPv4ワイルドカードを使用してデフォルトのポータルも作成しています。これは、任意のIPv4アドレスが設定されたターゲットにアクセスできることを意味しています。

次のステップは、iSCSIターゲットのLUN(論理ユニット番号)を作成することです。これを行う最適な方法は、targetcliでその名前と番号を自動的に割り当てることです。iSCSIターゲットのディレクトリに切り替えて、lunディレクトリのcreateコマンドを使用して、LUNをバックストアに割り当てます。

```

/> cd /iscsi/iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456/
/iscsi/iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456> cd tpg1
/iscsi/iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456/tpg1> luns/
create /backstores/fileio/test-disc

```

lsコマンドを実行して、変更を確認します。

```
/iscsi/iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456/tpg1> ls
```

```
o- tpg1 ..... [no-gen-acls, no-auth]
  o- acls ..... [ACLs: 0]
  o- luns ..... [LUNs: 1]
    | o- lun0 ..... [fileio/test-disc (/alt/test.img) (default_tg_pt_gp)]
    o- portals ..... [Portals: 1]
      o- 0.0.0.0:3260 ..... [OK]
```

これで現在は、1GBのファイルベースのバックストアを持つiSCSIターゲットが存在します。ターゲットには`iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456`という名前があり、システムの任意のネットワークポートからアクセスできます。

最後に、イニシエータが設定されたターゲットにアクセスできることを確認する必要があります。これを行う1つの方法は、各イニシエータに対してACLルールを作成し、ターゲットへの接続を許可することです。この場合、そのIQNを使用して必要な各イニシエータを一覧にする必要があります。既存のイニシエータのIQNは、`/etc/iscsi/initiatorname.iscsi`ファイルにあります。次のコマンドを使用して、必要なイニシエータ（この場合は、`iqn.1996-04.de.suse:01:54cab487975b`）を追加します。

```
/iscsi/iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456/tpg1> acls/ create
iqn.1996-04.de.suse:01:54cab487975b
Created Node ACL for iqn.1996-04.de.suse:01:54cab487975b
Created mapped LUN 0.
/iscsi/iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456/tpg1>
```

または、アクセス制限のないデモモードでターゲットを実行することもできます。この方法は安全性は低くなりますが、デモ目的や閉じたネットワークで実行する場合は役立つ可能性があります。デモモードを有効にするには、次のコマンドを実行します。

```
/iscsi/iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456/tpg1> set attribute
generate_node_acls=1
/iscsi/iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456/tpg1> set attribute
demo_mode_write_protect=0
```

最後のステップは、ルートディレクトリで使用可能な`saveconfig`コマンドを使用して、作成された設定を保存することです。

```
/> saveconfig /etc/target/example.json
```

保存されたファイルから設定を復元する必要がある場合は、まず現在の設定をクリアする必要があります。最初に設定を保存しない限り、現在の設定をクリアすると、データが失われることに注意してください。次のコマンドを使用して設定をクリアして再ロードします。

```
/> clearconfig
As a precaution, confirm=True needs to be set
/> clearconfig confirm=true
All configuration cleared
/> restoreconfig /etc/target/example.json
Configuration restored from /etc/target/example.json
```

/>

設定されたターゲットが機能しているかどうかテストするには、同じシステムにインストールされた `open-iscsi` iSCSIイニシエータを使用してそれに接続します(`HOSTNAME`をローカルマシンのホスト名で置き換えます)。

```
> iscsiadadm -m discovery -t st -p HOSTNAME
```

たとえば次のように、このコマンドは検出されたターゲットのリストを返します。

```
192.168.20.3:3260,1 iqn.2003-01.org.linux-iscsi.e83.x8664:sn.8b35d04dd456
```

その後 `login` iSCSIコマンドを使用して一覧表示されているターゲットに接続できます。これにより、ローカルディスクとしてターゲットが使用可能になります。

15.5 インストール時のiSCSIディスクの使用

iSCSI対応のファームウェアを使用している場合は、AMD64/Intel 64およびIBM POWERの各アーキテクチャ上のiSCSIディスクからのブートがサポートされています。

インストール時にiSCSIディスクを使用するには、次のパラメータをブートパラメータ行に追加する必要があります。

```
withiscsi=1
```

インストール中に、インストールプロセスで使用するiSCSIディスクをシステムに接続するオプションが記載された、追加の画面が表示されます。



注記: マウントポイントのサポート

iSCSIデバイスはブートプロセス中は非同期で表示されます。これらのデバイスがルートファイルシステム用に正しく設定されていることがinitrdによって保証されるまでの間、他のファイルシステムや`/usr`などのマウントポイントでは、これは保証されません。したがって、`/usr`や`/var`などのシステムマウントポイントはサポートされません。これらのデバイスを使用するには、各サービスとデバイスが正しく同期されていることを確認します。

15.6 iSCSIのトラブルシューティング

本項では、iSCSIターゲットとiSCSIイニシエータに関するいくつかの既知の問題と、考えられる解決策について説明します。

15.6.1 iSCSI LIOターゲットサーバにターゲットLUNをセットアップする際のポータルエラー

iSCSI LIOターゲットグループの追加または編集を行う際に、次のエラーが発生する:

```
Problem setting network portal IP_ADDRESS:3260
```

/var/log/YaST2/y2logログファイルに、次のエラーが含まれている:

```
find: `/sys/kernel/config/target/iscsi': No such file or directory
```

この問題は、iSCSI LIOターゲットサーバソフトウェアがその時点で実行中ではない場合に発生します。この問題を解決するには、YaSTを終了して、コマンドラインで手動で`systemctl start targetcli`を実行してiSCSI LIOを起動し、再試行します。

次のように入力して、`configfs`、`iscsi_target_mod`、および`target_core_mod`がロードされているかどうかをチェックすることもできます。サンプルの応答を示しています。

```
> sudo lsmod | grep iscsi
iscsi_target_mod      295015  0
target_core_mod       346745  4
iscsi_target_mod,target_core_pscsi,target_core_iblock,target_core_file
configfs              35817   3 iscsi_target_mod,target_core_mod
scsi_mod              231620  16
iscsi_target_mod,target_core_pscsi,target_core_mod,sg,sr_mod,mptctl,sd_mod,
scsi_dh_rdac,scsi_dh_emc,scsi_dh_alua,scsi_dh_hp_sw,scsi_dh,libata,mptspi,
mptscsih,scsi_transport_spi
```

15.6.2 iSCSI LIOターゲットが他のコンピュータで表示されない

ターゲットサーバでファイアウォールを使用している場合は、他のコンピュータでiSCSI LIOターゲットを表示できるようにするために使用するiSCSIポートを開く必要があります。詳細については、「[15.2.1項 「iSCSI LIOターゲットサービスの起動およびファイアウォールの設定」](#)」を参照してください。

15.6.3 iSCSIトラフィックのデータパッケージがドロップされる

ファイアウォールは、過剰にビジーになるとパケットをドロップすることがあります。SUSEファイアウォールのデフォルトは、3分後にパケットをドロップすることです。iSCSIトラフィックのパケットがドロップされていることが分かった場合は、ファイアウォールがビジーになったとき、パケットをドロップする代わりにキューに入れるように、SUSEファイアウォールを設定することを検討してください。

15.6.4 LVMでiSCSIボリュームを使用する

iSCSIターゲットでLVMを使用する際には、本項のトラブルシューティングのヒントを使用してください。

15.6.4.1 ブート時にiSCSIイニシエータの検出が行われるかどうかを確認する

iSCSIイニシエータをセットアップする際には、udevがブート時にiSCSIデバイスを検出し、LVMによるそれらのデバイスの使用をセットアップできるように、ブート時の検出を有効にしてください。

15.6.4.2 iSCSIターゲットの検出がブート時に起きることを確認する

udevは、デバイスのデフォルトセットアップを提供することを思い出してください。デバイスを作成するすべてのアプリケーションがブート時に起動されることを確認してください。これにより、udevがシステム起動時にそれらを認識し、デバイスを割り当てることができます。アプリケーションまたはサービスが後まで起動しない場合は、udevがブート時のように自動的にデバイスを作成することはありません。

15.6.5 設定ファイルが手動に設定されていると、iSCSIターゲットがマウントされる

Open-iSCSIは、/etc/iscsi/iscsid.confファイルでnode.startupオプションが手動に設定されている場合でも、設定ファイルを手動で変更すれば、起動時にターゲットをマウントできます。

/etc/iscsi/nodes/TARGET_NAME/IP_ADDRESS,PORT/defaultファイルを調べます。このファイルには、/etc/iscsi/iscsid.confファイルを上書きするnode.startup設定が含まれています。YaSTインターフェースを使用してマウントオプションを手動に設定すると、/etc/iscsi/nodes/TARGET_NAME/IP_ADDRESS,PORT/defaultファイルでもnode.startup = manualが設定されます。

15.7 iSCSI LIOターゲットの用語

backstore

iSCSIのエンドポイントの基礎となる実際のストレージを提供する、物理的ストレージオブジェクト。

CDB (command descriptor block)

SCSIコマンドの標準フォーマットCDBは一般的に6、10、または12バイトの長さですが、16バイトまたは可変長でも構いません。

CHAP (Challenge Handshake Authentication Protocol)

ポイントツーポイントプロトコル(PPP)の認証方法で、あるコンピュータのアイデンティティを別のコンピュータに対して確認するために使用します。Link Control Protocol (LCP)によって2台のコンピュータが接続され、CHAPメソッドがネゴシエートされた後、認証者はランダムなチャレンジをピアに送信します。ピアは、チャレンジおよび秘密鍵に依存した、暗号学的にハッシュされたレスポンスを発行します。認証者は、ハッシュされたレスポンスを、予想されるハッシュ値の自身の計算に対して検証し、認証を了承するか、接続を終了します。CHAPは、RFC 1994で定義されています。

CID (接続識別子)

イニシエータが生成する16ビットの番号で、2つのiSCSIデバイス間の接続を、一意に識別するもの。この番号は、ログインフェーズの間に提示されます。

エンドポイント

iSCSIターゲット名とiSCSI TPG (IQN + Tag)の組み合わせ

EUI (extended unique identifier)

世界中のあらゆるデバイスを一意に識別する、64ビットの番号。フォーマットは、会社ごとに一意である24ビットと、その会社が自社の各デバイスに割り当てる40ビットで構成されます。

イニシエータ

SCSIセッションの開始エンド。通常は、コンピュータなどの制御デバイス。

IPS (Internet Protocol storage)

IPプロトコルを使用してストレージネットワーク内のデータを移動する、プロトコルまたはデバイスのクラス。FCIP (Fibre Channel over Internet Protocol)、iFCP (Internet Fibre Channel Protocol)、およびiSCSI (Internet SCSI)は、すべてIPSプロトコルの例です。

IQN (iSCSI qualified name)

世界中のあらゆるデバイスを一意に識別する、iSCSIの名前形式(たとえば:
iqn.5886.com.acme.tapedrive.sn-a12345678)。

ISID (initiator session identifier)

イニシエータが生成する48ビットの番号で、イニシエータとターゲット間のセッションを一意に識別するもの。この値はログインプロセスの間に作成され、ログインPDUとともにターゲットに送られます。

MCS (multiple connections per session)

iSCSI仕様の一部で、イニシエータとターゲット間での複数のTCP/IP接続を可能にするもの。

MPIO (multipath I/O)

サーバとストレージ間でデータが複数の冗長パスをとることができるもの。

ネットワークポータル

iSCSIエンドポイントおよびIPアドレスとTCP(転送制御プロトコル)ポートの組み合わせ。TCPポート3260が、iSCSIプロトコル用のポート番号です。これは、IANA(Internet Assigned Numbers Authority)により定義されています。

SAM (SCSI architectural model)

SCSIの動作を一般的な表記で記載した文書で、異なる種類のデバイスがさまざまなメディア上で通信することを可能にするもの。

ターゲット

SCSIセッションの受信側で、通常はディスクドライブ、テープドライブ、スキャナなどのデバイス。

ターゲットグループ(TG)

ビューの作成時にすべて同じ扱いを受ける、SCSIターゲットポートのリスト。ビューを作成することで、LUN(論理ユニット番号)のマッピングが簡素化されます。それぞれのビューエントリが、ターゲットグループ、ホストグループ、およびLUNを指定します。

ターゲットポート

iSCSIエンドポイントと、1つ以上のLUNの組み合わせ。

ターゲットポートグループ(TPG)

IPアドレスとTCPポート番号のリストで、特定のiSCSIターゲットがどのインターフェースから受信するかを決定するもの。

ターゲットセッション識別子(TSID)

ターゲットが生成する 16ビットの番号で、イニシエータとターゲット間のセッションを一意に識別するもの。この値はログインプロセスの間に作成され、ログインレスポンス PDU(プロトコルデータユニット)とともにイニシエータに送られます。

15.8 詳細情報

iSCSIプロトコルは、数年に渡って利用されています。iSCSIとSANソリューションの比較やパフォーマンスのベンチマークは多くのレビューで取り上げられており、ハードウェアソリューションについて説明したドキュメントもあります。詳細については、<https://www.open-iscsi.com/>にあるOpen-iSCSIプロジェクトのホームページを参照してください。

また、[iscsiadm](#)、[iscsid](#)のマニュアルページ、および設定ファイルの例/[etc/iscsid.conf](#)も参照してください。

16 Fibre Channel Storage over Ethernet Networks: FCoE

多くの企業のデータセンターが、そのLANおよびデータトラフィックをEthernetに依存し、またそのストレージインフラストラクチャをファイバチャネルに依存しています。Open Fibre Channel over Ethernet (FCoE)イニシエータソフトウェアは、Ethernetアダプタが付いたサーバが、Ethernetネットワーク上でファイバチャネルストレージに接続できるようにします。このコネクティビティはこれまで、ファイバチャネルファブリック上にファイバチャネルアダプタを有するシステム用に、独占的に確保されていました。FCoEテクノロジは、ネットワークコンバージェンスを支援することで、データセンター内の複雑性を減らします。これにより、ファイバチャネルストレージへの既存の投資を無駄にすることなく、ネットワーク管理を簡素化することができます。

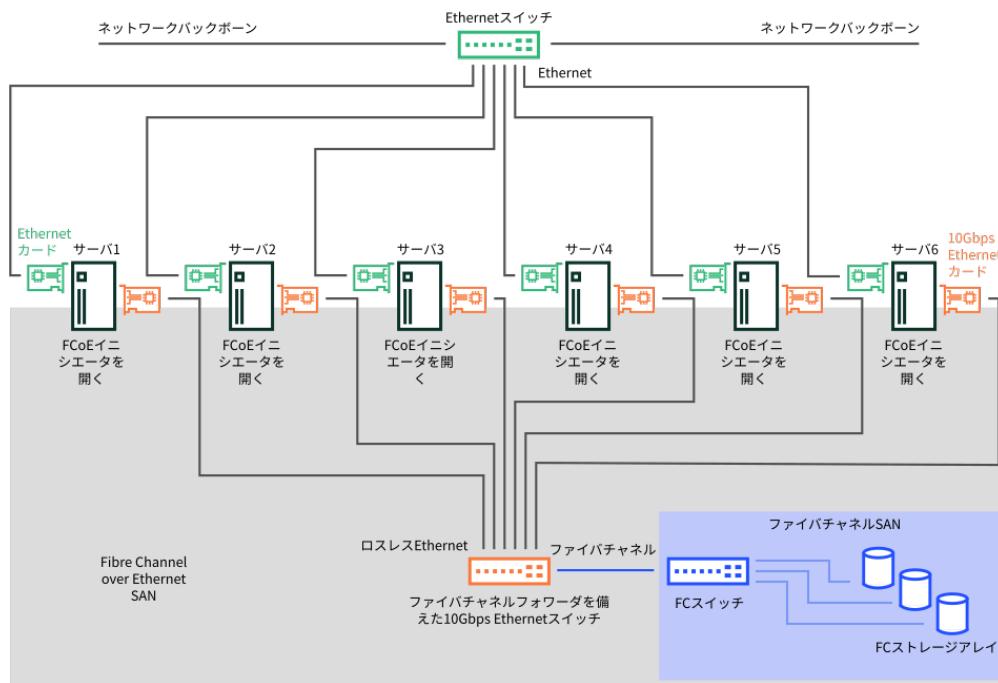


図 16.1: OPEN FIBRE CHANNEL OVER ETHERNET SAN

Open-FCoEでは、ホストバスアダプタ上の専有のハードウェアではなく、ホストでファイバチャネルのプロトコルを実行することができます。対象としているのは10 Gbps (ギガバイト/秒)のEthernetアダプタですが、PAUSEフレームに対応したすべてのEthernetアダプタで使用可能です。イニシエータソフトウェアにより、ファイバチャネルプロトコルの処理モジュールと、Ethernetベースのトランスポートモジュールが提供されます。Open-FCoEモ

ジユールは、SCSI用の低レベルドライバの役割を果たします。Open-FCoEトランスポートは、net_deviceを使用してパケットの送受信を行います。DCB(データセンターブリッジング)ドライバにより、FCoE向けのサービスの質が提供されます。

FCoEは、ファイバチャネルフレームを変えずに、ファイバチャネルのプロトコルトラフィックをEthernet接続上で動かす、カプセル化プロトコルです。これにより、ネットワークセキュリティとトラフィック管理インフラストラクチャが、ファイバチャネルにおけるのと同じようにFCoEでも機能することができます。

以下の条件が当てはまる企業では、FCoEの導入を選択してもよいでしょう。

- すでにファイバチャネルストレージシステムがあり、ファイバチャネルのスキルと知識を持つ管理者がいる。
- ネットワーク内に、10 GbpsのEthernetを展開している。

本項では、ネットワークにFCoEを設定する方法を説明します。

16.1 インストール時におけるFCoEインタフェースの設定

SUSE Linux Enterprise Server向けのYaSTのインストールでは、サーバとファイバチャネルストレージインフラストラクチャ間の接続用のスイッチでFCoEが有効になっていれば、オペレーティングシステムのインストール時にFCoEディスクの設定を行うことができます。一部のシステムBIOSタイプでは、FCoEディスクを自動的に検出することができ、そのディスクをYaSTのインストールソフトウェアに報告します。ただし、FCoEディスクの自動検出は、すべてのBIOSのタイプでサポートされているわけではありません。その場合、インストールの開始時に次のwithfcoeオプションをカーネルのコマンドラインに追加することで、自動検出を有効にすることができます。

```
withfcoe=1
```

FCoEディスクが検出されると、YaSTのインストールでは、同時にFCoEを設定するオプションがあります。[ディスクアクティベーション] ページで、FCoEインターフェースの設定を選択して、FCoEの設定にアクセスします。FCoEインターフェースの設定については、16.3項「[YaSTを使用したFCoEサービスの管理](#)」を参照してください。



注記: マウントポイントのサポート

FCoEデバイスはブートプロセス中は非同期で表示されます。これらのデバイスがルートファイルシステム用に正しく設定されていることがinitrdによって保証されるまでの間、他のファイルシステムや`/usr`などのマウントポイントでは、これは保証されません。したがって、`/usr`や`/var`などのシステムマウントポイントはサポートされません。これらのデバイスを使用するには、各サービスとデバイスが正しく同期されていることを確認します。

16.2 FCoEおよびYaSTのFCoEクライアントのインストール

サーバへの接続用のスイッチでFCoEを有効にすることで、ストレージインフラストラクチャ内にFCoEディスクを設定することができます。SUSE Linux Enterprise Serverオペレーティングシステムのインストール時にFCoEディスクが利用可能であれば、FCoEイニシエータソフトウェアが、その時点で自動的にインストールされます。

FCoEイニシエータソフトウェアとYaST FCoEクライアントソフトウェアがインストールされていない場合は、次の手順で次のコマンドを使用して手動でインストールします。

```
> sudo zypper in yast2-fcoe-client fcoe-utils
```

または、YaSTソフトウェアマネージャを使用して、これらのパッケージをインストールします。

16.3 YaSTを使用したFCoEサービスの管理

YaST FCoEクライアント設定オプションを使用して、お使いのファイバチャネルストレージインターフェース内にFCoEディスク用のFCoEインタフェースの作成、設定、および削除ができます。このオプションを使用するには、FCoEイニシエータサービス(fcoemonデーモン)およびLink Layer Discovery Protocolエージェントデーモン(lldpad)がインストールされて実行中であり、FCoE接続が、FCoE対応のスイッチで有効になっている必要があります。

1. YaSTを起動し、ネットワークサービス › FCoEクライアントの設定の順に選択します。

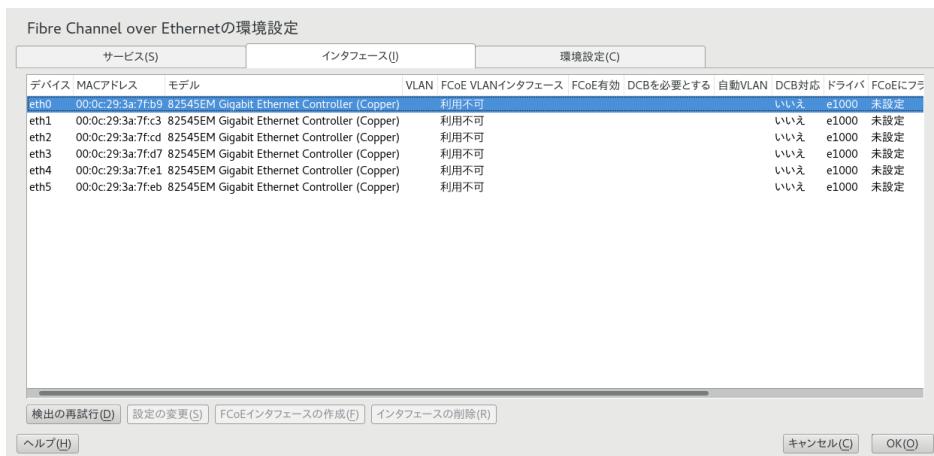


2. サービスタブで、FCoEサービスとLldpad (Link Layer Discovery Protocolエージェントデーモン)サービスの開始時刻を確認し、必要に応じて変更します。

- FCoEサービスの開始:** Fibre Channel over Ethernetサービスのfcoemonデーモンを、サーバの起動時に開始するか、マニュアルで開始するかを指定します。このデーモンは、FCoEインターフェースを制御して、llpadデーモンとの接続を確立します。値は、起動時(デフォルト)またはマニュアルです。
- Lldpadサービスの開始:** Link Layer Discovery Protocolエージェントllpadデーモンを、サーバの起動時に開始するか、マニュアルで開始するかを指定します。llpadデーモンは、データセンター bridging 機能およびFCoEインターフェースの設定について、fcoemonデーモンに情報を送ります。値は、起動時(デフォルト)またはマニュアルです。

設定を変更した場合は、OKをクリックして変更内容を保存して適用します。

3. インタフェースタブで、サーバ上で検出されたすべてのネットワークアダプタに関する情報(VLANおよびFCoEの設定に関する情報を含む)を確認します。また、FCoE VLANインターフェースの作成や既存のFCoEインターフェース設定の変更、FCoEインターフェースの削除もできます。



FCoE VLANインターフェース列を使用して、FCoEが使用可能かどうかを判断します。

Interface Name

インターフェースに名前が割り当てられている(eth4.200など)場合は、スイッチでFCoEが利用可能であり、FCoEインターフェースがアダプタに対してアクティブになっています。

設定されていません:

状態が未設定である場合は、スイッチでFCoEが有効になっていますが、FCoEインターフェースはアダプタに対してアクティブになっていません。アダプタでインターフェースを有効にするには、アダプタを選択して、FCoE VLANインターフェースを作成をクリックします。

使用不可:

状態が使用不可である場合は、FCoEがスイッチ上のその接続に対して有効になつていませんため、そのアダプタではFCoEは使えません。

4. 未設定のFCoE対応アダプタを設定するには、そのアダプタを選択し、FCoE VLANインターフェースを作成をクリックします。問い合わせに対して、はいを選択して確認します。

アダプタがインターフェース名と共にFCoE VLANインターフェース列に表示されます。

5. 設定済みのアダプタの設定を変更するには、リストでそのアダプタを選択し、Change Settings (設定の変更)をクリックします。

次のオプションを設定できます。

FCoEの有効化

アダプタに対してFCoEインスタンスの作成を有効または無効にします。

DCBが必要

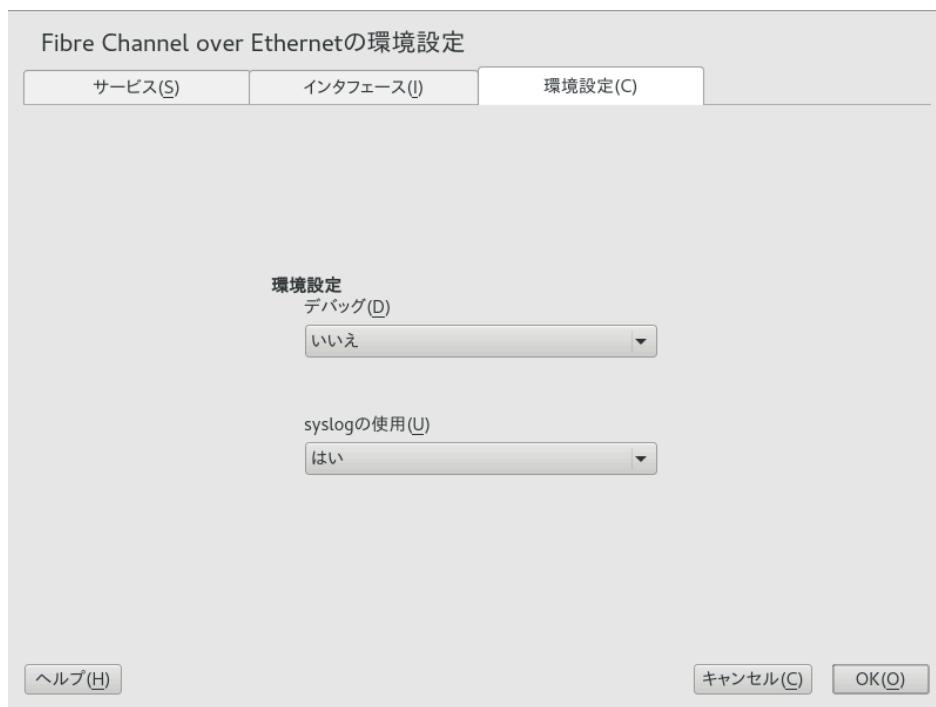
データセンターブリッジングがアダプタに必要かどうかを指定します(通常は必要です)。

Auto VLAN

fcoemonデーモンでVLANインターフェースを作成するかどうかを指定します。

設定を変更した場合は、次へをクリックして変更内容を保存して適用します。設定は、/etc/fcoe/cfg-ethXファイルに書き込まれます。fcoemonデーモンは、初期化時に各FCoEインターフェースの環境設定ファイルを読み込みます。

6. 設定済みのインターフェースを削除するには、それをリストで選択します。インターフェースの削除をクリックし、続行をクリックして確認します。FCoEインターフェースの値が、未構成に変わります。
7. 設定タブで、FCoEシステムサービスの全般設定を確認または変更します。FCoEサービススクリプトとfcoemonデーモンからのデバッグメッセージを有効/無効にしたり、メッセージをシステムログに送信するかどうかを指定したりできます。



8. OKをクリックして、変更内容を保存して適用します。

16.4 コマンドを使用したFCoEの設定

次の各ステップでは、fipvlanコマンドを使用する必要があります。このコマンドがインストールされていない場合、次のコマンドを実行してインストールします。

```
> sudo zypper in fcoe-utils
```

すべてのEthernetインターフェースを検出し、設定するには、次の手順に従います。

1. ターミナルを開きます。
2. 使用可能なすべてのEthernetインターフェースを検出するには、次のコマンドを実行します。

```
> sudo fipvlan -a
```

3. FCoEオフロードが設定されている各Ethernetインターフェースに対して次のコマンドを実行します。

```
> sudo fipvlan -c -s ETHERNET_INTERFACE
```

このコマンドは、ネットワークインターフェースが存在しない場合はそれを作成し、検出されたFCoE VLANでOpen-FCoEイニシエータを起動します。

16.5 FCoE管理ツールを使用したFCoEインスタンスの管理

fcoeadmユーティリティは、FCoE (Fibre Channel over Ethernet)管理ツールです。これを使って、所定のネットワークインターフェースのFCoEインスタンスの作成、破棄、およびリセットを行うことができます。**fcoeadm**ユーティリティは、ソケットインターフェースを通じて、実行中の**fcoemon**プロセスにコマンドを送ります。**fcoemon**の詳細については、[man 8 fcoemon](#)を参照してください。

fcoeadmユーティリティを使用して、以下に関してFCoEインスタンスにクエリを行うことができます。

- インタフェース
- ターゲットLUN
- ポートの統計データ

fcoeadmユーティリティは、[fcoe-utils](#)パッケージの一部です。このコマンドの一般的な構文は、次のようにになります。

```
fcoeadm
  [-c|--create] [<ethX>]
  [-d|--destroy] [<ethX>]
  [-r|--reset] [<ethX>]
  [-S|--Scan] [<ethX>]
  [-i|--interface] [<ethX>]
  [-t|--target] [<ethX>]
  [-l|--lun] [<ethX>]
  [-s|--stats <ethX>] [<interval>]
  [-v|--version]
  [-h|--help]
```

詳細については、[man 8 fcoeadm](#)を参照してください。

例

```
fcoeadm -c eth2.101
```

FCoEインスタンスをeth2.101上に作成します。

fcoeadm -d eth2.101

FCoEインスタンス上のeth2.101を破棄します。

fcoeadm -i eth3

インターフェースeth3上のFCoEインスタンスすべてに関する情報を表示します。インターフェースが指定されていない場合、FCoEインスタンスが作成されているすべてのインターフェースの情報を表示します。次に、接続eth0.201の情報の例を示します。

```
> sudo fcoeadm -i eth0.201
Description:      82599EB 10-Gigabit SFI/SFP+ Network Connection
Revision:        01
Manufacturer:    Intel Corporation
Serial Number:   001B219B258C
Driver:          ixgbe 3.3.8-k2
Number of Ports: 1

Symbolic Name:   fcoe v0.1 over eth0.201
OS Device Name: host8
Node Name:       0x1000001B219B258E
Port Name:       0x2000001B219B258E
FabricName:     0x2001000573D38141
Speed:          10 Gbit
Supported Speed: 10 Gbit
MaxFrameSize:   2112
FC-ID (Port ID): 0x790003
State:          Online
```

fcoeadm -l eth3.101

接続eth3.101で検出されたすべてのLUNの詳細情報を表示します。接続が指定されていない場合、すべてのFCoE接続で検出されたすべてのLUNの情報を表示します。

fcoeadm -r eth2.101

eth2.101上のFCoEインスタンスをリセットします。

fcoeadm -s eth3 3

FCoEインスタンスが存在する特定のeth3ポートに関する統計情報を3秒間隔で表示します。統計情報は、時間間隔ごとに1行ずつ表示されます。間隔を指定していない場合、デフォルトの1秒が間隔として使用されます。

fcoeadm -t eth3

FCoEインスタンスが存在する特定のeth3ポートから検出されたすべてのターゲットに関する情報を表示します。検出された各ターゲットの後ろに、関連付けられたLUNが列記されます。インスタンスが指定されていない場合、FCoEインスタンスが存在するすべてのポートからのターゲットを表示します。次に、接続eth0.201からのターゲットの情報の例を示します。

```

> sudo fcoeadm -t eth0.201
  Interface:      eth0.201
  Roles:          FCP Target
  Node Name:     0x200000D0231B5C72
  Port Name:     0x210000D0231B5C72
  Target ID:      0
  MaxFrameSize:   2048
  OS Device Name: rport-8:0-7
  FC-ID (Port ID): 0x79000C
  State:          Online

  LUN ID  Device Name  Capacity  Block Size  Description
  -----  -----
    40  /dev/sdqi     792.84 GB    512        IFT DS S24F-R2840-4 (rev 386C)
    72  /dev/sdpk     650.00 GB    512        IFT DS S24F-R2840-4 (rev 386C)
   168  /dev/sdgy     1.30 TB      512        IFT DS S24F-R2840-4 (rev 386C)

```

16.6 詳細情報

詳細については、以下のマニュアルを参照してください。

- Open-FCoEのサービスデーモンについては、[fcoemon\(8\)](#)マニュアルページを参照してください。
- Open-FCoEの管理ツールについては、[fcoeadm\(8\)](#)マニュアルページを参照してください。
- データセンターブリッジング設定ツールについては、[dcbtool\(8\)](#)マニュアルページを参照してください。
- Link Layer Discovery Protocolエージェントデーモンについては、[lldpad\(8\)](#)マニュアルページを参照してください。

17 NVMe-oF

この章では、NVMe over Fabricホストおよびターゲットの設定方法について説明します。

17.1 概要

NVM Express® (NVMe®)は、不揮発性ストレージ(通常はSSDディスク)にアクセスするためのインターフェース規格です。NVMeはSATAをはるかに上回る処理速度をサポートし、レイテンシも低くなります。

NVMe-oF™はRDMA、TCP、**NVMe over Fibre Channel (FC-NVMe)**などの異なるネットワーキングファブリックを介してNVMeストレージにアクセスするためのアーキテクチャです。NVMe-oFの役割はiSCSIと同様です。耐障害性を向上させるため、NVMe-oFにはマルチパスのサポートが組み込まれています。NVMe-oFマルチパスは、従来のデバイスマッパー・マルチパスに基づいていません。

NVMeホストは、NVMeターゲットに接続するマシンです。NVMeターゲットは、そのNVMeブロックデバイスを共有するマシンです。

NVMeはSUSE Linux Enterprise Server 15 SP6でサポートされています。NVMeブロックストレージおよびNVMe-oFターゲットとホストには、専用のカーネルモジュールが用意されています。

ご使用のハードウェアに関する特別な考慮事項があるかどうかを確認するには、[17.4項 「特定のハードウェアの設定」](#)を参照してください。

17.2 NVMe-oFホストの設定

NVMe-oFを使用するには、サポートされているネットワーキング方法のいずれかでターゲットを使用可能にする必要があります。NVMe over Fibre Channel、TCP、およびRDMAがサポートされています。以降のセクションでは、NVMe over FabricホストをNVMeターゲットに接続する方法について説明します。

17.2.1 コマンドラインクライアントのインストール

NVMe-oFを使用するには、`nvme`コマンドラインツールが必要です。インストールするには、`zypper`を実行します。

```
> sudo zypper in nvme-cli
```

すべての使用可能なサブコマンドを一覧にするには、`nvme --help`を使用します。`nvme`サブコマンド用のマニュアルページが提供されています。`man nvme-SUBCOMMAND`を実行すると、このページを参照できます。たとえば、`discover`サブコマンドのマニュアルページを参照するには、`man nvme-discover`を実行します。

17.2.2 NVMe-oFターゲットの検出

NVMe-oFターゲットで使用可能なNVMeサブシステムを一覧にするには、検出コントローラのアドレスとサービスIDが必要です。

```
> sudo nvme discover -t TRANSPORT -a DISCOVERY_CONTROLLER_ADDRESS -s SERVICE_ID
```

`TRANSPORT`は、基盤となる転送メディア(`loop`、`rdma`、`tcp`、または`fc`)で置き換えます。`DISCOVERY_CONTROLLER_ADDRESS`は、検出コントローラのアドレスで置き換えます。RDMAとTCPの場合、これはIPv4アドレスである必要があります。`SERVICE_ID`は、転送サービスIDで置き換えます。RDMAまたはTCPのように、サービスがIPベースの場合、サービスIDはポート番号を指定します。ファイバチャネルの場合、サービスIDは必要ありません。

NVMeホストは、接続が許可されているサブシステムのみを参照します。

例:

```
> sudo nvme discover -t tcp -a 10.0.0.3 -s 4420
```

FCの場合、次のようにになります。

```
> sudo nvme discover --transport=fc \
    --traddr=nn-0x201700a09890f5bf:pn-0x201900a09890f5bf \
    --host-traddr=nn-0x200000109b579ef6:pn-0x100000109b579ef6
```

詳細については、`man nvme-discover`を参照してください。

17.2.3 NVMe-oFターゲットへの接続

NVMeサブシステムを特定した後で、`nvme connect`コマンドを使用して接続できます。

```
> sudo nvme connect -t transport -a DISCOVERY_CONTROLLER_ADDRESS -s SERVICE_ID -n SUBSYSTEM_NQN
```

`TRANSPORT`は、基盤となる転送メディア(`loop`、`rdma`、`tcp`または`fc`)で置き換えます。`DISCOVERY_CONTROLLER_ADDRESS`は、検出コントローラのアドレスで置き換えます。RDMAとTCPの場合、これはIPv4アドレスである必要があります。`SERVICE_ID`は、転送

サービスIDで置き換えます。RDMAまたはTCPのように、サービスがIPベースの場合、これはポート番号を指定します。SUBSYSTEM_NQNは、検出コマンドによって検出された、目的のサブシステムのNVMe修飾名で置き換えます。NQNは、NVMe修飾名(NVMe Qualified Name)の略語です。NQNは固有である必要があります。

例:

```
> sudo nvme connect -t tcp -a 10.0.0.3 -s 4420 -n nqn.2014-08.com.example:nvme:nvm-subsystem-sn-d78432
```

FCの場合、次のようにになります。

```
> sudo nvme connect --transport=fc \
    --traddr=nn-0x201700a09890f5bf:pn-0x201900a09890f5bf \
    --host-traddr=nn-0x200000109b579ef6:pn-0x100000109b579ef6 \
    --nqn=nqn.2014-08.org.nvmeexpress:uuid:1a9e23dd-466e-45ca-9f43-a29aaf47cb21
```

または、nvme connect-allを使用して、すべての検出されたネームスペースに接続します。高度な使用法については、man nvme-connectおよびman nvme-connect-allを参照してください。

パスが失われると、NVMeサブシステムでは、nvme connectコマンドのctrl-loss-tmoオプションで定義された時間、再接続しようとします。この時間(デフォルト値は600秒)が経過したら、パスが削除され、ブロックレイヤの上位レイヤ(ファイルシステム)に通知されます。デフォルトでは、ファイルシステムは、読み取り専用にマウントされます。これは通常望ましい動作ではありません。したがって、NVMeサブシステムが制限なしで再接続を試行し続けるようにctrl-loss-tmoオプションを設定することをお勧めします。そのためには、次のコマンドを実行します。

```
> sudo nvme connect --ctrl-loss-tmo=-1
```

NVMe over Fabricsサブシステムを起動時に使用できるようにするには、ホストに/etc/nvme/discovery.confファイルを作成し、パラメータをdiscoverコマンドに渡します([17.2.2項「NVMe-oFターゲットの検出」](#)を参照)。たとえば、次のようにdiscoverコマンドを使用します。

```
> sudo nvme discover -t tcp -a 10.0.0.3 -s 4420
```

discoverコマンドのパラメータを/etc/nvme/discovery.confファイルに追加します。

```
echo "-t tcp -a 10.0.0.3 -s 4420" | sudo tee -a /etc/nvme/discovery.conf
```

次に、nvmf-autoconnectサービスを有効にします。

```
> sudo systemctl enable nvmf-autoconnect.service
```

17.2.4 マルチパス処理

NVMeネイティブマルチパス処理はデフォルトで有効になっています。コントローラID設定のCMICオプションが設定されている場合、NVMeスタックはNVMEドライブをデフォルトでマルチパスデバイスとして認識します。

マルチパスを管理するには、以下を使用できます。

マルチパスの管理

nvme list-subsy

マルチパスデバイスのレイアウトを印刷します。

multipath -ll

コマンドには互換性モードがあり、NVMeマルチパスデバイスを表示します。enable_foreignオプションを有効にしてこのコマンドを使用する必要があることに留意してください。詳細については、[18.13項 「その他のオプション」](#) を参照してください。

nvme-core.multipath=N

オプションがブートパラメータとして追加されると、NVMeネイティブマルチパスが無効になります。

17.3 NVMe-oFターゲットの設定

17.3.1 コマンドラインクライアントのインストール

NVMe-oFターゲットを設定するには、nvmetcliコマンドラインツールが必要です。インストールするには、zypperを実行します。

```
> sudo zypper in nvmetcli
```

nvmetcliの現在のドキュメントはhttps://git.infradead.org/users/hch/nvmetcli.git/blob_plain/HEAD:/Documentation/nvmetcli.txtから入手できます。

17.3.2 設定手順

次の手順に、NVMe-oFターゲットの設定方法の例を示します。

設定はツリー構造で格納されます。移動するには、`cd`コマンドを使用します。オブジェクトを一覧にするには、`ls`を使用します。`create`を使用して新しいオブジェクトを作成できます。

1. `nvmetcli`インタラクティブシェルを起動します。

```
> sudo nvmetcli
```

2. 新しいポートを作成します。

```
(nvmetcli)> cd ports  
(nvmetcli)> create 1  
(nvmetcli)> ls 1/  
o- 1  
  o- referrals  
  o- subsystems
```

3. NVMeサブシステムを作成します。

```
(nvmetcli)> cd /subsystems  
(nvmetcli)> create nqn.2014-08.org.nvmexpress:NVMf:uuid:c36f2c23-354d-416c-95de-f2b8ec353a82  
(nvmetcli)> cd nqn.2014-08.org.nvmexpress:NVMf:uuid:c36f2c23-354d-416c-95de-f2b8ec353a82/  
(nvmetcli)> ls  
o- nqn.2014-08.org.nvmexpress:NVMf:uuid:c36f2c23-354d-416c-95de-f2b8ec353a82  
  o- allowed_hosts  
  o- namespaces
```

4. 新しいネームスペースを作成し、そのネームスペースにNVMeデバイスを設定します。

```
(nvmetcli)> cd namespaces  
(nvmetcli)> create 1  
(nvmetcli)> cd 1  
(nvmetcli)> set device path=/dev/nvme0n1  
Parameter path is now '/dev/nvme0n1'.
```

5. 以前に作成したネームスペースを有効にします。

```
(nvmetcli)> cd ..  
(nvmetcli)> enable  
The Namespace has been enabled.
```

6. 作成したネームスペースを表示します。

```
(nvmetcli)> cd ..  
(nvmetcli)> ls  
o- nqn.2014-08.org.nvmexpress:NVMf:uuid:c36f2c23-354d-416c-95de-f2b8ec353a82  
  o- allowed_hosts
```

```
o- namespaces  
  o- 1
```

7. すべてのホストがサブシステムを使用できるようにします。この操作は、セキュリティ保護された環境でのみ実行します。

```
(nvmetcli)> set attr allow_any_host=1  
Parameter allow_any_host is now '1'.
```

または、特定のホストのみが接続できるようにします。

```
(nvmetcli)> cd nqn.2014-08.org.nvmeexpress:NVMf:uuid:c36f2c23-354d-416c-95de-f2b8ec353a82/allowed_hosts/  
(nvmetcli)> create hostnqn
```

8. すべての作成されたオブジェクトを一覧にします。

```
(nvmetcli)> cd /  
(nvmetcli)> ls  
o- /  
  o- hosts  
  o- ports  
    | o- 1  
    |   o- referrals  
    |   o- subsystems  
  o- subsystems  
    o- nqn.2014-08.org.nvmeexpress:NVMf:uuid:c36f2c23-354d-416c-95de-f2b8ec353a82  
      o- allowed_hosts  
      o- namespaces  
        o- 1
```

9. TCPを介してターゲットを使用できるようにします。RDMAにはtrtype=rdmaを使用します。

```
(nvmetcli)> cd ports/1/  
(nvmetcli)> set addr adrfam=ipv4 trtype=tcp traddr=10.0.0.3 trsvcid=4420  
Parameter trtype is now 'tcp'.  
Parameter adrfam is now 'ipv4'.  
Parameter trsvcid is now '4420'.  
Parameter traddr is now '10.0.0.3'.
```

または、ファイバチャネルを介して使用可能にすることができます。

```
(nvmetcli)> cd ports/1/  
(nvmetcli)> set addr adrfam=fc trtype=fc  
  traddr=nn-0x1000000044001123:pn-0x2000000055001123 trsvcid=none
```

10. サブシステムをポートにリンクします。

```
(nvmetcli)> cd /ports/1/subsystems  
(nvmetcli)> create nqn.2014-08.org.nvmeexpress:NVMf:uuid:c36f2c23-354d-416c-95de-f2b8ec353a82
```

dmesgを使用してポートが有効になっていることを確認できるようになりました。

```
# dmesg  
...  
[ 257.872084] nvmet_tcp: enabling port 1 (10.0.0.3:4420)
```

17.3.3 ターゲット設定のバックアップと復元

次のコマンドを使用してJSONファイルにターゲット設定を保存できます。

```
> sudo nvmetcli  
(nvmetcli)> saveconfig nvme-target-backup.json
```

設定を復元するには、次のコマンドを使用します。

```
(nvmetcli)> restore nvme-target-backup.json
```

現在の設定を消去することもできます。

```
(nvmetcli)> clear
```

17.4 特定のハードウェアの設定

17.4.1 概要

一部のハードウェアでは、正しく動作させるために特殊な設定が必要です。次の各セクションの見出しを参照し、記載されているデバイスまたはベンダのいずれかに該当しないか確認してください。

17.4.2 Broadcom

Broadcom Emulex LightPulse Fibre Channel SCSI ドライバを使用している場合は、lpfcモジュールのターゲットおよびホスト上にカーネル設定パラメータを追加します。

```
> sudo echo "options lpfc lpfc_enable_fc4_type=3" > /etc/modprobe.d/lpfc.conf
```

Broadcomアダプタファームウェアのバージョンが11.4.204.33以降であることを確認します。現在のバージョンのnvme-cli、nvmetcli、およびカーネルがインストールされていることも確認します。

ファイバチャネルポートをNVMeターゲットとして有効にするには、追加のモジュールパラメータを設定する必要があります。たとえば、lpfc_enable_nvmet=COMMA_SEPARATED_WWPNSと指定します。先行する0xとともにWWPNを入力します。たとえば、lpfc_enable_nvmet=0x2000000055001122,0x2000000055003344と指定します。一覧表示されているWWPNのみがターゲットモードに設定されます。ファイバチャネルポートは、ターゲットまたはイニシエータとして設定できます。

17.4.3 Marvell

FC-NVMeは、QLE269xおよびQLE27xxアダプタでサポートされています。FC-NVMeのサポートは、Marvell® QLogic® QLA2xxxファイバチャネルドライバでデフォルトで有効になっています。

NVMeが有効になっていることを確認するには、次のコマンドを実行します。

```
> cat /sys/module/qla2xxx/parameters/ql2xnvmeenable
```

結果の1は、NVMeが有効になっていることを示し、0は無効になっていることを示します。次に、Marvellアダプタファームウェアが少なくともバージョン8.08.204であることを次のコマンドの出力をチェックして確認します。

```
> cat /sys/class/scsi_host/host0/fw_version
```

最後に、SUSE Linux Enterprise Serverで使用可能な最新バージョンのnvme-cli、QConvergeConsoleCLI、およびカーネルがインストールされていることを確認します。たとえば、次を実行して

```
# zypper lu && zypper pchk
```

更新とパッチを確認します。

インストールに関する詳細については、次のMarvellユーザガイドのFC-NVMeのセクションを参照してください。

- https://driverdownloads.qlogic.com/QLogicDriverDownloads_UI>ShowEula.aspx?resourceid=32769&docid=96728&ProductCategory=39&Product=1259&Os=126 ↗
- https://driverdownloads.qlogic.com/QLogicDriverDownloads_UI>ShowEula.aspx?resourceid=32761&docid=96726&ProductCategory=39&Product=1261&Os=126 ↗

17.5 NVMe-oF over TCPからのブート

SLESでは、NVM Express® Boot Specification 1.0 (<https://nvmexpress.org/wp-content/uploads/NVM-Express-Boot-Specification-2022.11.15-Ratified.pdf>) に従って、NVMe-oF over TCPからのブートをサポートしています。

リモートストレージサーバへのNVMe-oF over TCP接続を試行し、これらの接続をブートに使用するように、UEFIプリブート環境を設定できます。プリブート環境では、ACPIテーブル (NVMe Boot Firmware Table (NBFT))を作成し、ブートに使用されるNVMe-oF設定に関する情報を格納します。オペレーティングシステムは後のブートステージでこのテーブルを使用し、ネットワーキングとNVMe-oF接続を設定して、ルートファイルシステムにアクセスします。

17.5.1 システム要件

NVMe-oF over TCPからシステムをブートするには、次の要件を満たしている必要があります。

- SLES15 SP6以降。
- NVMe-oF over TCPをサポートするSANストレージアレイ。
- NVMe-oF over TCPからのブートをサポートするBIOSを備えたホストシステム。この機能のサポートについては、ハードウェアのベンダにお問い合わせください。NVMe-oF over TCPからのブートは現在、UEFIプラットフォームでのみサポートされています。

17.5.2 インストール

NVMe-oF over TCPからSLESをインストールするには、次の手順に従います。

1. ホストシステムのUEFIセットアップメニューを使用して、ブート時にNVMe-oF接続が確立されるように設定します。通常、ネットワーキング(ローカルIPアドレス、ゲートウェイなど)とNVMe-oFターゲット(リモートIPアドレス、サブシステムNQN、または検出NQN)の両方を設定する必要があります。設定の説明については、ハードウェアのマニュアルを参照してください。ハードウェアベンダが、BIOS設定を一元的かつリモートで管理する手段を提供している場合があります。詳細については、ハードウェアベンダに問い合わせてください。
2. 『展開ガイド』の説明に従ってインストールを準備します。

3. サポートされているインストール方法を使用して、システムのインストールを開始します。NVMe-oF over TCPへのインストールを有効にするために、特定のブートパラメータを使用する必要はありません。
4. BIOSが正しく設定されている場合は、YaSTのディスクパーティション設定ダイアログに、BIOSで設定されたサブシステムによってエクスポートされたNVMeネームスペースが表示されます。これらはNVMeデバイスとして表示され、`tcp`文字列は、デバイスがTCPトランSPORTを介して接続されていることを示しています。これらのネームスペースにオペレーティングシステム(特にEFIブートパーティションとルートファイルシステム)をインストールします。
5. インストールを完了します。

インストール後、システムはNVMe-oF over TCPから自動的にブートするはずです。ブートしない場合は、ブートの優先度がBIOSセットアップで正しく設定されているか確認します。

ブートに使用されるネットワークインターフェース名は、`nbft0`、`nbft1`などです。NVMe-oFブートに関する情報を取得するには、次のコマンドを実行します。

```
# nvme nbft show
```

17.6 詳細情報

`nvme`コマンドの機能の詳細については、`nvme nvme-help`を参照してください。

次のリンクには、NVMeおよびNVMe-oFの概要があります。

- <https://nvmexpress.org/> ↗
- https://www.nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf ↗
- <https://storpool.com/blog/demystifying-what-is-nvmeof> ↗

18 デバイスのマルチパスI/Oの管理

本項では、マルチパスI/O (MPIO)を使用して、サーバ/ブロックストレージデバイス間のマルチパスのフェールオーバーおよびパスの負荷分散を管理する方法について説明します。

18.1 マルチパスI/Oの理解

マルチパス処理とは、サーバのホストバスアダプタおよびデバイスのストレージコントローラ間で、複数の物理パスをまたいで、同じ物理または論理ブロックストレージデバイスと通信するサーバの機能です。これは、通常、FC (Fibre Channel)環境またはiSCSI SAN環境で行われます。

Linuxマルチ処理は、接続に耐障害性を与え、アクティブな接続全体に負荷を分散します。マルチパス処理が設定および実行されると、自動的に、デバイス接続の障害が特定され、I/Oが代替の接続に再経路指定されます。

マルチパス処理は、接続の障害に対して耐障害性を提供しますが、ストレージデバイス自体の障害に対する耐障害性は提供しません。後者は、ミラーリングのような補完テクニックによって提供されます。

18.1.1 マルチパスの用語

ストレージアレイ

SANストレージをクライアントに提供する多数のディスクおよび複数のファブリック接続(コントローラ)を備えたハードウェアデバイス。ストレージアレイは通常、RAIDとフェールオーバー機能を備えており、マルチパス処理をサポートしています。これまで、アクティブ/パッシブ(フェールオーバー)とアクティブ/アクティブ(負荷分散)のストレージアレイ設定は区別されていました。これらの概念は今でも存在しますが、最新のハードウェアでサポートされるパスグループとアクセス状態の概念の特殊なケースにすぎません。

ホスト、ホストシステム

ストレージアレイのクライアントシステムとして動作するSUSE Linux Enterprise Serverを実行しているコンピュータ。

マルチパスマップ、マルチパスデバイス

一連のパスデバイスです。ストレージアレイ上のストレージボリュームを表し、ホストシステムからは単一のブロックデバイスと認識されます。

パスデバイス

マルチパスマップのメンバー(通常はSCSIデバイス)です。各パスデバイスは、ホストコンピュータと実際のストレージボリューム(iSCSIセッションの論理ユニットなど)との間の固有の接続を表します。

WWID

「World Wide Identifier」です。[multipath-tools](#)ではWWIDを使用して、マルチパスマップにアセンブルする必要がある低レベルデバイスを判断します。WWIDは、設定可能なマップ名と区別する必要があります([18.12項 「マルチパスデバイス名とWWID」](#)を参照)。

uevent、udevイベント

カーネルによってユーザースペースに送信され、[udev](#)サブシステムによって処理されるイベント。デバイスの追加や削除、またはプロパティの変更を行うと、ueventが生成されます。

デバイスマッパー

仮想ブロックデバイスを作成するためのLinuxカーネルのフレームワークです。マップデバイスに対するI/O操作は基礎となるブロックデバイスにリダイレクトされます。デバイスマッピングはスタックされる場合があります。デバイスマッパーでは独自のイベントシグナル処理を実装します。これは「デバイスマッパーイベント」または「dmイベント」とも呼ばれます。

initramfs

初期RAMファイルシステムは、これまでの経緯もあって、「初期RAMディスク」(initrd)とも呼びられます(『管理ガイド』、第16章「ブートプロセスの概要」、16.1項「用語集」を参照)。

ALUA

「Asymmetric Logical Unit Access」です。これはSCSI標準のSCSI-3で導入された概念です。ストレージボリュームには、さまざまな状態(アクティブ、スタンバイなど)でポートグループに編成されている複数のポートからアクセスできます。ALUAは、ポートグループとその状態を問い合わせ、ポートグループの状態を変更するためのSCSIコマンドを定義します。SCSIをサポートする最新のストレージアレイは通常ALUAもサポートします。

18.2 ハードウェアサポート

マルチパスドライバとツールは、SUSE Linux Enterprise Serverでサポートされているすべてのアーキテクチャで利用できます。プロトコルを区別しない汎用ドライバは、市販のほとんどのマルチパス対応ストレージハードウェアで動作します。一部のストレージアレイベンダは、独自のマルチパス処理管理ツールを提供しています。ベンダのハードウェアマニュアルを参照して、どのような設定が必要か判別してください。

18.2.1 マルチパスの実装: デバイスマッパーとNVMe

Linuxにおけるマルチパス処理で従来の一般的な実装では、デバイスマッパーフレームワークを使用します。SCSIデバイスなどのほとんどのデバイスタイプでは、デバイスマッパーのマルチパス処理が唯一使用可能な実装です。デバイスマッパーのマルチパスは、高度に設定可能で、柔軟性に優れています。

Linux **NVMe Express** (NVMe)カーネルサブシステムでは、カーネルでマルチパス処理をネイティブに実装します。通常高速で遅延が非常に小さいNVMeデバイスの演算オーバーヘッドが、この実装で軽減されています。ネイティブNVMeマルチパス処理ではユーザースペースコンポーネントは不要です。SLE 15以降、ネイティブマルチパス処理がNVMeマルチパスデバイスのデフォルトになっています。詳細については、[17.2.4項 「マルチパス処理」](#)を参照してください。

この章では、デバイスマッパーマルチパスおよびそのユーザースペースコンポーネント[multipath-tools](#)について説明します。[multipath-tools](#)は、ネイティブのNVMeマルチパス処理も限定的にサポートしています([18.13項 「その他のオプション」](#)を参照)。

18.2.2 マルチパス処理のストレージアレイ自動検出

デバイスマッパーのマルチパスは一般的な技術です。マルチパスデバイスの検出で必要なことは、低レベルデバイス(SCSIなど)がカーネルによって検出されることと、デバイスのプロパティが複数の低レベルデバイスを(実際に異なるデバイスではなく)同じボリュームへの異なる「パス」として確実に識別することのみです。

[multipath-tools](#)パッケージは、ベンダと製品名によってストレージアレイを検出します。これには、広範なストレージ製品に対してデフォルト設定が組み込まれています。ご使用のストレージアレイのハードウェアドキュメントを参照してください。Linuxのマルチパス処理設定に対して独自の推奨事項を提示しているベンダもあります。

使用しているストレージアレイの組み込み設定に変更を適用する必要がある場合、[18.8項 「マルチパス設定」](#)を参照してください。



重要: 組み込みのハードウェアプロパティに関する免責条項

multipath-toolsには、多くのストレージアレイ用の事前設定が組み込まれています。あるストレージ製品用にこのような事前設定が存在することは、そのストレージ製品のベンダがdm-multipathで製品をテストしたことを意味しませんし、ベンダがその製品でdm-multipathの使用に関して保証やサポートを行うことも「意味しません」。サポート関連の質問については、ベンダのドキュメントを必ず参照してください。

18.2.3 特定のハードウェアハンドラを必要とするストレージアレイ

あるパスから別のパスにフェールオーバーするための特殊なコマンドや標準と異なるエラー処理方法が必要なストレージアレイもあります。これらの特殊なコマンドと方法は、Linuxカーネルのハードウェアハンドラによって実装されています。最新のSCSIストレージアレイは、SCSI標準で定義されている「Asymmetric Logical Unit Access」(ALUA)ハードウェアハンドラをサポートしています。ALUAの他に、SLEカーネルには、Netapp E-Series(RDAC)、Dell/EMC CLARiiON CXアレイファミリ、およびHPのレガシアレイ用のハードウェアハンドラが含まれています。

Linuxカーネル4.4以降では、ほとんどのアレイ(ALUAをサポートするすべてのアレイを含む)のハードウェアハンドラをLinuxカーネルで自動検出します。要件は、それぞれのデバイスが検出されるときにデバイスハンドラモジュールがロードされることだけです。この要件は、multipath-toolsパッケージが適切な設定ファイルをインストールすることによって実現します。デバイスハンドラは、特定のデバイスにアタッチされると、変更できなくなります。

18.3 マルチパス処理のプラニング

マルチパスI/Oソリューションのプラニング時には、本項のガイドラインに従ってください。

18.3.1 前提条件

- マルチパス処理対象のデバイスに使用するストレージアレイで、マルチパス処理がサポートされている必要があります。詳細については、[18.2項 「ハードウェアサポート」](#) を参照してください。
- サーバのホストバスアダプタおよびブロックストレージデバイスのバスコントローラ間に複数の物理パスが存在している場合のみ、マルチパス処理を設定する必要があります。
- 一部のストレージアレイについては、アレイの物理および論理デバイスのマルチパス処理を管理するための独自のマルチパス処理ソフトウェアがベンダから提供されます。この場合は、ベンダの指示に従って、それらのデバイスのマルチ処理を設定してください。
- 仮想化環境でマルチパス処理を使用する場合、マルチパス処理は、ホストサーバ環境で制御されます。デバイスのマルチパス処理を設定してから、デバイスを仮想ゲストマシンに割り当ててください。

18.3.2 マルチパスのインストールタイプ

インストールタイプは、ルートデバイスを処理する方法で区別されます。[18.4項 「マルチパスシステムでのSUSE Linux Enterprise Serverのインストール」](#) では、インストール中およびインストール後にさまざまなセットアップを作成する方法について説明しています。

18.3.2.1 マルチパスのルートファイルシステム(SANブート)

ルートファイルシステムはマルチパスデバイス上にあります。これは通常、ディスクのないサーバでSANストレージのみを使用する場合です。このようなシステムでは、起動にマルチパスのサポートが必要であり、initramfsでマルチパス処理を有効にする必要があります。

18.3.2.2 ローカルディスクのルートファイルシステム

ルートファイルシステム(および場合によっては他のファイルシステム)は、ローカルストレージ(直接アタッチされたSATAディスクやローカルRAIDなど)上にありますが、このシステムではさらにマルチパスのSANストレージのファイルシステムを使用します。このシステムタイプは次の3つの方法で設定できます。

ローカルディスク用のマルチパスセットアップ

すべてのブロックデバイスはローカルディスクを含むマルチパスマップの一部です。ルートデバイスは、パスが1つしかないデイグレードマルチパスマップとして表示されます。YaSTによる初期システムインストール中にマルチパス処理が有効になると、この設定が作成されます。

ローカルディスクをマルチパスから除外する

この設定では、マルチパス処理はinitramfsで有効になっていますが、ルートデバイスはマルチパスから明示的に除外されます([18.11.1項 「multipath.confのblacklistセクション」を参照](#))。手順[18.1 「インストール後にルートディスクのマルチパス処理を無効にする」](#)では、この設定のセットアップ方法を説明しています。

initramfsでマルチパスを無効にする

YaSTによる初期システムインストール中にマルチパス処理が有効にならないと、このセットアップが作成されます。この設定はかなり脆弱です。代わりに、他のオプションのいずれかを使用することを検討してください。

18.3.3 ディスク管理タスク

サードパーティSANアレイ管理ツールやご使用のストレージアレイのユーザインタフェースを使用して、論理デバイスを作成し、それらをホストに割り当てます。両側でホストの資格情報を正しく設定してください。

実行中のホストにボリュームを追加したり、削除したりできますが、変更を検出するには、SCSIターゲットを再スキャンし、ホスト上でマルチパス処理を再設定する必要がある場合があります。[18.14.6項 「新規デバイスのスキャン\(再起動なし\)」](#)を参照してください。



注記: ストレージプロセッサ

一部のディスクアレイでは、ストレージアレイがストレージプロセッサを使用してトラフィックを管理します。1つのプロセッサがアクティブとなり、もう1つのプロセッサは障害が発生するまでパッシブとなります。パッシブストレージプロセッサに接続している場合、目的のLUNが表示されないか、表示はされるがアクセスしようとするとI/Oエラーが発生する場合があります。

ディスクアレイに複数のストレージプロセッサがある場合は、アクセスしたいLUNを所有するアクティブなストレージプロセッサにSANスイッチが接続していることを必ず確認してください。

18.3.4 ソフトウェアRAIDと複雑なストレージスタック

マルチパス処理は、SCSIデバイスなどの基本的なストレージデバイスの上にセットアップされます。マルチレイヤのストレージスタックでは、マルチパス処理は常に最下位レイヤです。ソフトウェアRAID、論理ボリューム管理、ブロックデバイスの暗号化などのその他のレイヤは、マルチパス処理の上に重ねられます。したがって、複数のI/Oパスを持ち、ソフトウェアRAIDで使用予定の各デバイスは、まず、マルチパス処理用に設定してから、ソフトウェアRAIDデバイスとして作成する必要があります。

18.3.5 高可用性ソリューション

ストレージリソースのクラスタリング用の高可用性ソリューションは、各ノード上でマルチパス処理サービスをベースとして実行されます。各ノード上の`/etc/multipath.conf`ファイル内の構成設定が、クラスタ全体で同一であるようにしてください。

マルチパスデバイスがすべてのデバイス間で同じ名前であるようにしてください。詳細については、[18.12項 「マルチパスデバイス名とWWID」](#) を参照してください。

LAN上のデバイスをミラーリングするDRBD (Distributed Replicated Block Device)高可用性ソリューションは、マルチパス処理をベースとして実行されます。複数のI/Oパスを持ち、DRDBソリューションで使用予定のデバイスごとに、マルチパス処理用デバイスを設定してから、DRBDを設定する必要があります。

`pacemaker`と`sbd`などフェンシングに共有ストレージを使用するクラスタリングソフトウェアと共にマルチパス処理を使用する場合、特別な注意が必要です。詳細については[18.9.2項 「クラスタ化されたサーバでのキューポリシー」](#) を参照してください。

18.4 マルチパスシステムでのSUSE Linux Enterprise Serverのインストール

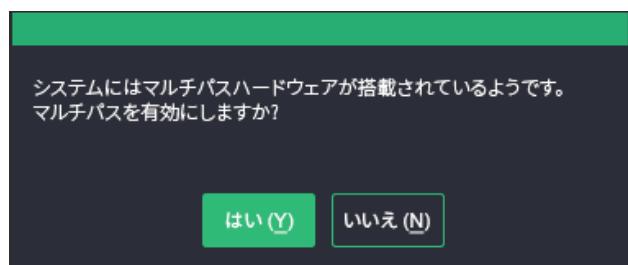
マルチパスハードウェアを含むシステムでSUSE Linux Enterprise Serverをインストールするために必要な特別なインストールパラメータはありません。

18.4.1 接続されているマルチパスデバイスを使用しないインストール

後で、システムにマルチパスSANデバイスを追加する目的で、最初にFabricとストレージを設定せずに、ローカルディスク上でインストールを実行したい場合があります。この場合、インストールは、非マルチパスシステム上でのインストールのように続行します。インストール後、`multipath-tools`がインストールされますが、`systemd`サービス(`multipathd.service`)は無効になります。システムは、[18.3.2.2項 「ローカルディスクのルートファイルシステム」](#) の`initramfs`でマルチパスを無効にするで説明しているように設定されます。SANハードウェアを追加する前に、`multipathd.service`を有効にし、開始する必要があります。ルートデバイス用に`/etc/multipath.conf`に`blacklist`エントリを作成することをお勧めします ([18.11.1項 「multipath.confのblacklistセクション」](#) を参照)。

18.4.2 接続されているマルチパスデバイスによるインストール

マルチパスデバイスが、インストール時にシステムに接続される場合、YaSTはそれらを検出し、パーティショニングステージに入る前にマルチパスを有効にするかどうかを尋ねるポップアップウィンドウを表示します。



このプロンプトで「いいえ」を選択すると(非推奨)、インストールは[18.4.1項 「接続されているマルチパスデバイスを使用しないインストール」](#)のように続行します。パーティショニングステージでは、後でマルチパスマップの一部になるデバイスを使用/編集しないでください。マルチパスのプロンプトで「はい」を選択すると、インストール中に`multipathd`が実行されます。`/etc/multipath.conf`の`blacklist`セクションにはデバイスは追加されないため、パーティショニングダイアログでは、ローカルディスクを含むすべてのSCSIデバイスとDASDデバイスがマルチパスデバイスとして表示されます。インストール後、[18.3.2.1項 「マルチパスのルートファイルシステム\(SANブート\)」](#)で説明されているように、すべてのSCSIデバイスとDASDデバイスがマルチパスデバイスになります。

手順 18.1: インストール後にルートディスクのマルチパス処理を無効にする

この手順は、ローカルディスクにインストールし、インストール中にマルチパス処理を有効にしていることを前提としているため、ルートデバイスがマルチパス上にありますが、[18.3.2.2項 「ローカルディスクのルートファイルシステム」のローカルディスクをマルチパスから除外する](#)で説明されているように、システムを設定したいと考えています。

1. システムをチェックしてローカルルートデバイスへの`/dev/mapper/...`参照を探し、それらをデバイスがマルチパスマップでない場合にも機能する参照で置き換えます([18.12.4項 「マルチパスマップの参照」](#)を参照)。次のコマンドを実行しても参照が見つからない場合、変更を適用する必要はありません。

```
> sudo grep -rl /dev/mapper/ /etc
```

2. `dracut`の`by-uuid`永続デバイスピリシーに切り替えます([18.7.4.2項 「initramfsの永続的なデバイス名」](#)を参照)。

```
> echo 'persistent_policy="by-uuid"' | \
    sudo tee /etc/dracut.conf.d/10-persistent-policy.conf
```

3. ルートデバイスのWWIDを決定します。

```
> multipathd show paths format "%i %d %w %s"
0:2:0:0 sda 3600605b009e7ed501f0e45370aaeb77f IBM,ServeRAID M5210
...
```

このコマンドは、すべてのパスデバイスとそのWWID、およびベンダ/製品情報を出力します。ルートデバイス(ここでは、ServeRAIDデバイス)を識別し、WWIDをメモすることができます。

4. 決定したWWIDを使用して`/etc/multipath.conf` ([18.11.1項 「multipath.confのblacklistセクション」](#)を参照)にブラックリストエントリを作成します(まだこれらの設定を適用しないでください)。

```
blacklist {
    wwid 3600605b009e7ed501f0e45370aaeb77f
}
```

5. initramfsを再構築します。

```
> sudo dracut -f
```

6. 再起動します。非マルチパスルートディスクを使用してシステムが起動します。

18.5 マルチパスシステムでのSLEの更新

システムをオンラインで更新する場合、『アップグレードガイド』、第5章「オンラインでのアップグレード」の説明に従って続行できます。

システムのオフライン更新は、[18.4項 「マルチパスシステムでのSUSE Linux Enterprise Serverのインストール」](#)で説明されている新規インストールと似ています。`blacklist`がないため、ユーザがマルチパスを有効にすると、ルートデバイスがマルチパスデバイスとして表示されます(これがマルチパスデバイスでない場合も同様)。更新手順中に`dracut`が`initramfs`を構築するとき、ブートシステムに表示されるストレージスタックとは異なるストレージスタックが表示されます。[18.7.4.2項 「initramfsの永続的なデバイス名」](#)および[18.12.4項 「マルチパスマップの参照」](#)を参照してください。

18.6 マルチパス管理ツール

SUSE Linux Enterprise Serverのマルチパス処理のサポートは、Linuxカーネルのデバイスマッパー・マルチパスモジュールと[multipath-tools](#)ユーザースペースパッケージに基づいています。

一般的なマルチパス処理機能は、デバイスマッパーのマルチパス(DM-MP)モジュールによって処理されます。詳細については、[18.6.1項 「デバイスマッパー・マルチパスモジュール」](#)を参照してください。

[multipath-tools](#)および[kpartx](#)パッケージは、自動パス検出とグループ化を処理するツールを提供します。ツールを次に示します。

multipathd

マルチパスマップを設定および監視するデーモン、およびデーモンプロセスと通信するコマンドラインクライアント。[18.6.2項 「multipathdデーモン」](#)を参照してください。

multipath

マルチパス操作用のコマンドラインツール。[18.6.3項 「multipathコマンド」](#)を参照してください。

kpartx

マルチパスデバイスの「パーティション」を管理するためのコマンドラインツール。[18.7.3項 「マルチパスデバイスのパーティションとkpartx」](#)を参照してください。

mpathpersist

SCSIの永続的な予約を管理するためのコマンドラインツール。[18.6.4項 「SCSIの永続的な予約とmpathpersist」](#)を参照してください。

18.6.1 デバイスマッパーマルチパスモジュール

デバイスマッパーマルチパス(DM-MP)モジュール(`dm-multipath.ko`)は、Linuxに一般的なマルチパス処理機能を提供します。DM-MPIOは、SCSIおよびDASDデバイスに対してSUSE Linux Enterprise Serverでマルチパス処理を行う際に推奨されるソリューションであり、NVMeデバイスにも使用できます。



注記: NVMeデバイスに対してDM-MPを使用する

SUSE Linux Enterprise Server 15以降、ネイティブNVMeマルチパス処理(18.2.1項「マルチパスの実装: デバイスマッパーとNVMe」を参照)がNVMeに推奨され、デフォルトで使用されています。ネイティブNVMeマルチパス処理を無効にし、代わりにデバイスマッパーマルチパスを使用するには(「非推奨」)、カーネルパラメータ`nvme-core.multipath=0`を指定してブートします。

デバイスマッパーマルチパスモジュールは次のタスクを処理します。

- アクティブパスグループ内の複数のパスに負荷を分散する。
- パスデバイスのI/Oエラーを通知し、これらを障害とマークして、I/Oがこれらに送信されないようにする。
- アクティブパスグループのすべてのパスで障害が発生したときにパスグループを切り替える。
- すべてのパスで障害が発生した場合、設定に応じてマルチパスデバイスのI/Oを失敗させるまたはI/Oをキューに入れる。

次のタスクは、デバイスマッパーのマルチパスモジュールではなく、`multipath-tools`パッケージのユーザースペースコンポーネントによって処理されます。

- 同じストレージデバイスへのさまざまなパスを表すデバイスを検出し、それからマルチパスマップをアセンブルする。
- 似ているプロパティを持つパスデバイスをパスグループに収集する。
- パスデバイスをアクティブに監視して、障害または再インスタンス化を探す。
- パスデバイスの追加と削除を監視する。
- デバイスマッパーマルチパスモジュールは、セットアップと設定に使用しやすいユーザインターフェースを提供しません。

`multipath-tools`パッケージのコンポーネントの詳細については、18.6.2項

「`multipathd`デーモン」を参照してください。



注記: マルチパスが防ぐ障害

DM-MPIOは、メディアエラーなどのデバイス自体の障害ではなく、デバイスへのパスの障害からシステムを保護します。デバイス自体の障害は、レプリケーションなど別の方法で防ぐ必要があります。

18.6.2 `multipathd` デーモン

`multipathd` は、最新Linuxデバイスマッパーのマルチパスにおけるセットアップの最重要部分です。これは通常、systemdサービス`multipathd.service`を通じて開始されます(18.7.1項「マルチパスサービスの有効化、起動、および停止」を参照)。

`multipathd` は次のタスクを実行します(設定によって異なるものもあります)。

- 起動時に、パスデバイスを検出し、検出されたデバイスからマルチパスマップを設定します。
- ueventとデバイスマッパーイベントを監視し、必要に応じてマルチパスマップでパスマッピングの追加や削除を行い、フェールオーバーまたはフェールバック操作を開始します。
- 新しいパスデバイスが検出されるとすぐに新しいマップをセットアップします。
- パスデバイスを定期的にチェックして障害を検出し、障害が発生したパスをテストして、正常に戻ったら復帰させます。
- すべてのパスで障害が発生した場合、`multipathd` はそのマップを無効にするか、またはマップデバイスを指定時間でキュー待ちモードに切り替えます。
- パスの状態の変更を処理し、必要に応じて、パスグループを切り替えたり、パスを再グループ化したりします。
- パスが「ぎりぎりの」状態(すなわち、動作状態と非動作状態の間でパスの状態が切り替わる不安定なファブリック状態)であるかどうかをテストします。
- 設定されている場合、パスデバイスでSCSIの永続的な予約キーを処理します。18.6.4項「SCSIの永続的な予約と`mpathpersist`」を参照してください。

`multipathd` では、コマンドラインのクライアントとしても動作し、インタラクティブコマンドを実行デーモンに送信することでコマンドを処理します。デーモンにコマンドを送信する一般的な構文は次のとおりです。

```
> sudo multipathd COMMAND
```

あるいは、

```
> sudo multipathd -k 'COMMAND'
```

複数の後続のコマンドを送信できる対話的モードもあります。

```
> sudo multipathd -k
```

注記: multipathとmultipathdを同時に動作させる方法

多くのmultipathdのコマンドには、multipathの同等コマンドがあります。たとえば、multipathd show topologyの動作はmultipath -llの動作と同じです。顕著な違いは、multipathdコマンドは実行中のmultipathdデーモンの内部状態を問い合わせるのに対して、multipathはカーネルおよびI/O操作から直接情報を取得することです。

マルチパスデーモンが実行中である場合、multipathdコマンドを使用してシステムを変更することをお勧めします。このようにしないと、デーモンが設定の変更に気づき、変更に反応する場合があります。場合によっては、適用された変更をデーモンで元に戻そうとします。multipathは、実行中のデーモンが検出された場合、マップの破棄やフラッシュなどの特定の危険性のあるコマンドをmultipathdに自動的に委任します。

下記のリストでは、使用頻度が高いmultipathdコマンドについて説明します。

show topology

現在のマップトポロジとプロパティを表示します。[18.14.2項 「マルチパスI/Oステータスの解釈」](#)を参照してください。

show paths

現在既知のパスデバイスを表示します。

show paths format "FORMAT STRING"

フォーマット文字列を使用して現在既知のパスデバイスを表示します。サポートされているフォーマット指定子のリストを表示するには、show wildcardsを使用します。

show maps

現在設定されているマップデバイスを表示します。

show maps format FORMAT STRING

フォーマット文字列を使用して、現在設定されているマップデバイスを表示します。サポートされているフォーマット指定子のリストを表示するには、show wildcardsを使用します。

show config local

`multipathd`が使用している現在の設定を表示します。

reconfigure

設定ファイルを再読み込みし、デバイスを再スキャンして、マップを再度セットアップします。これは`multipathd`の再起動と基本的に同じです。いくつかのオプションは再起動しないと変更できません。これらについてはマニュアルページの`multipath.conf(5)`で説明します。`reconfigure`コマンドを実行すると、何らかの方法で変更されたマップデバイスのみが再ロードされます。すべてのマップデバイスの再ロードを強制実行するには、`reconfigure all`を使用します(SUSE Linux Enterprise Server 15 SP4以降で使用できます。以前のバージョンでは`reconfigure`を使用してすべてのマップを再ロードしていました)。

del map MAP DEVICE NAME

指定されたマップデバイスとそのパーティションを設定解除し、削除します。`MAP` `DEVICE NAME`には、`dm-0`などのデバイスノード名、WWID、またはマップ名を使用できます。このコマンドは、デバイスを使用中には失敗します。

switchgroup map MAP DEVICE NAME group N

指定の数値インデックスを持つパスグループに切り替えます(先頭は1)。これは、手動フェールバックを使用するマップに対して有用です([18.9項 「フェールオーバー、キューリング、およびフェールバックのポリシーの設定」](#)を参照)。

パス状態の変更、キューの有効化または無効化などを実行できるコマンドも利用できます。詳細については`multipathd(8)`を参照してください。

18.6.3 `multipath`コマンド

マルチパスのセットアップはほとんど自動化されており、`multipathd`で処理されますが、`multipath`も一部の管理タスクで依然として役立ちます。このコマンドの使用例を次に示します。

`multipath`

パスデバイスを検出し、検出したすべてのマルチパスマップを設定します。

`multipath -d`

`multipath`に似ていますが、マップをセットアップしません(試行動作)。

multipath DEVICENAME

特定のマルチパスデバイスを設定します。DEVICENAMEは、デバイスノード名(`/dev/sdb`)またはデバイス番号(`major:minor`フォーマット)でメンバーパスデバイスを示すことができます。または、WWIDやマルチパスマップの名前も使用できます。

multipath -f DEVICENAME

マルチパスマップとそのパーティションマッピングを設定解除(「フラッシュ」)します。そのパーティションのいずれかまたはマップが使用中の場合、このコマンドは失敗します。DEVICENAMEで使用できる値については上記を参照してください。

multipath -F

すべてのマルチパスマップとそのパーティションマッピングを設定解除(「フラッシュ」)します。マップを使用中の場合、このコマンドは失敗します。

multipath -ll

現在設定されているすべてのマルチパスデバイスのステータスとトポロジを表示します。[18.14.2項 「マルチパスI/Oステータスの解釈」](#)を参照してください。

multipath -ll DEVICENAME

指定されたマルチパスデバイスのステータスを表示します。DEVICENAMEで使用できる値については上記を参照してください。

multipath -t

マルチパスの内部ハードウェアテーブルとアクティブな設定を表示します。設定パラメータの詳細については、[`multipath.conf\(5\)`](#)を参照してください。

multipath -T

[`multipath -t`](#)コマンドの機能と似ていますが、ホストで検出されたハードウェアと一致するハードウェアエントリのみを表示します。

`-v`オプションによって、出力の詳細レベルが制御されます。指定した値によって、[`/etc/multipath.conf`](#)のverbosityオプションが上書きされます。[18.13項 「その他のオプション」](#)を参照してください。

18.6.4 SCSIの永続的な予約とmpathpersist

mpathpersistユーティリティを使用して、デバイスマッパーマルチパスのデバイスでSCSIの永続的な予約を管理します。永続的な予約を行うと、SCSIの論理ユニットへのアクセスが特定のSCSIイニシエータに制限されます。マルチパス設定では、指定ボリュームですべてのI_T関連付け(パス)に同じ予約キーを使用することが重要です。そのようにしないと、あるパスデバイスで予約を作成すると別のパスでI/Oエラーが発生する場合があります。

このユーティリティを`/etc/multipath.conf`ファイルの`reservation_key`属性と共に使用して、SCSIデバイスの永続的な予約を設定します。このオプションが設定されている場合(のみ)、`multipathd`デーモンは新しく検出したパスまたは復帰したパスの永続的な予約をチェックします。

この属性は、`multipath.conf`の`defaults`セクションまたは`multipaths`セクションに追加できます。例:

```
multipaths {
    multipath {
        wwid          3600140508dbcf02acb448188d73ec97d
        alias         yellow
        reservation_key 0x123abc
    }
}
```

永続的な管理に適用可能なすべてのmpathデバイスに対して`reservation_key`パラメータを設定した後、`multipathd reconfigure`を使用して設定を再ロードします。



注記: 使用 “`reservation_key file`”

特別な値である`reservation_key file`が`multipath.conf`の`defaults`セクションで使用される場合、`mpathpersist`を使用して`/etc/multipath/prkeys`ファイルで予約キーを動的に管理できます。

これは、マルチパスマップで永続的な予約を処理する際のお勧めの方法です。この方法はSUSE Linux Enterprise Server 12 SP4から使用できます。

mpathpersistコマンドを使用して、SCSIデバイスで構成されるマルチパスマップの永続的な予約を問い合わせて、設定します。詳細については、`mpathpersist(8)`のマニュアルページを参照してください。このコマンドラインオプションは、`sg3_utils`パッケージの`sg_persist`のオプションと同じです。`sg_persist(8)`のマニュアルページでは、このオプションの意味を詳細に説明しています。

次の例では、`DEVICE`は`/dev/mapper/mpatha`など、デバイスマッパーのマルチパスデバイスを示しています。下記のコマンドは、読みやすくするために長いオプションと共にリストしています。すべてのオプションには、`mpathpersist -oGS 123abc DEVICE`のように1文字の置換文字が含まれています。

mpathpersist --in --read-keys `DEVICE`

デバイスに登録されている予約キーを読み取ります。

mpathpersist --in --read-reservation `DEVICE`

デバイスの既存の予約を表示します。

mpathpersist --out --register --param-sark=123abc `DEVICE`

デバイスの予約キーを登録します。これにより、ホストですべてのI_T関連付け(パスデバイス)の予約キーが追加されます。

mpathpersist --out --reserve --param-rk=123abc --prout-type=5 `DEVICE`

以前登録したキーを使用して、デバイスのタイプ5(登録者のみの排他書き込み)の予約を作成します。

mpathpersist --out --release --param-rk=123abc --prout-type=5 `DEVICE`

デバイスのタイプ5の予約を解放します。

mpathpersist --out --register-ignore --param-sark=0 `DEVICE`

以前存在していた予約キーをデバイスから削除します。

18.7 マルチパス処理用システムの設定

18.7.1 マルチパスサービスの有効化、起動、および停止

マルチパスサービスを有効にしてブート時に起動するには、次のコマンドを実行します。

```
> sudo systemctl enable multipathd
```

実行中のシステムでサービスを手動で開始するには、次のように入力します。

```
> sudo systemctl start multipathd
```

サービスを再開するには、次のように入力します。

```
> sudo systemctl restart multipathd
```

ほとんどの状況で、サービスの再開は不要です。単にmultipathdに設定を再ロードさせるには、次のコマンドを実行します。

```
> sudo systemctl reload multipathd
```

サービスのステータスを確認するには、次のように入力します。

```
> sudo systemctl status multipathd
```

現行セッションのマルチパスサービスを停止するには、次のコマンドを実行します。

```
> sudo systemctl stop multipathd multipathd.socket
```

サービスを停止しても既存のマルチパスマップは削除されません。未使用マップを削除するには、次のコマンドを実行します。

```
> sudo multipath -F
```

 **警告:** `multipathd.service`を常に有効にしておいてください
マルチパスハードウェアを備えたシステムでは、`multipathd.service`を常に有効にして実行しておくことを強くお勧めします。このサービスは、`systemd`のソケットアクトイベーションメカニズムをサポートしていますが、このメカニズムに依存することはお勧めしません。このサービスが無効になっていると、マルチパスマップは起動時にセットアップされません。



注記: マルチパスの無効化

上記の警告にもかかわらずマルチパスを無効にする必要がある場合(サードパーティのマルチパス処理ソフトウェアを導入するなど)、次のように進めます。マルチパスデバイスへのハードコード化された参照をシステムで使用していないことを確認します([18.15.2項 「デバイス参照の問題の理解」](#)を参照)。

1回のシステムブートに対してのみマルチパス処理を無効にするには、カーネルパラメータの`multipath=off`を使用します。これはブートシステムとinitramfsの両方に影響します。この場合はinitramfsを再構築する必要はありません。

`multipathd`サービスを恒久的に無効にして、今後のシステムブートでこのサービスが開始されないようにするには、次のコマンドを実行します。

```
> sudo systemctl disable multipathd multipathd.socket
```

```
> sudo dracut --force --omit multipath
```

(マルチパスサービスを無効または有効にするときには必ずinitramfsを再構築してください。詳細については、[18.7.4項「initramfsの同期状態の維持」](#)を参照してください。)

「multipathを手動で実行するときにも」マルチパスデバイスが設定されないようにする場合は、initramfsを再構築する前に、/etc/multipath.confの最後に次の行を追加します。

```
blacklist {  
    wwid .*  
}
```

18.7.2 マルチパス処理用SANデバイスの準備

SANデバイスのマルチパスI/Oを設定する前に、必要に応じて、次のようにSANデバイスを準備してください。

- ベンダのツールで、SANデバイスを設定し、ゾーン化します。
- ベンダのツールで、ストレージアレイ上のホストLUNのパーミッションを設定します。
- SUSE Linux Enterprise Serverでホストバスアダプタ(HBA)用ドライバが同梱されていない場合、HBAベンダからLinuxドライバをインストールします。詳細については、ベンダの特定マニュアルを参照してください。

マルチパスデバイスが検出され、multipathd.serviceが有効になっている場合、マルチパスマップは自動的に作成されるはずです。作成されない場合、[18.15.3項「緊急モードでのトラブルシューティング手順」](#)に、状況の調査に使用できるシェルコマンドのリストが表示されます。LUNがHBAドライバによって認識されない場合は、SANのゾーン化セットアップをチェックします。特に、LUNのマスキングがアクティブであるかどうか、LUNがサーバに正しく割り当てられているかどうかをチェックしてください。

LUNがHBAドライバによって認識できるが、対応するブロックデバイスが作成されない場合は、追加のカーネルパラメータが必要な場合があります。SUSE KnowledgeのTID 3955167: Troubleshooting SCSI (LUN) Scanning Issues (<https://www.suse.com/support/kb/doc.php?id=3955167>)を参照してください。

18.7.3 マルチパスデバイスのパーティションとkpartx

マルチパスマップには、そのパスデバイスのようなパーティションを設けることができます。パーティションテーブルのスキャンとパーティションのデバイスノードの作成は、ユーザスペースで **kpartx**ツールによって実行されます。**kpartx**は、udevルールによって自動的に起動します。通常、手動での実行は不要です。マルチパスのパーティションを参照する方法については、[18.12.4項 「マルチパスマップの参照」](#)を参照してください。



注記: kpartxの起動の無効化

`/etc/multipath.conf`の`skip_kpartx`オプションを使用して、選択したマルチパスマップでの**kpartx**の起動を無効にできます。たとえば、これは仮想化ホストで有用である場合があります。

マルチパスデバイス上のパーティションテーブルとパーティションは、YaSTまたは**fdisk**や**parted**などのツールを使用して、通常どおり操作できます。パーティションテーブルに適用する変更は、パーティション処理ツールが終了するとシステムによって記録されます。これが動作しない場合(デバイスがビジーであることが原因であることが多い)、**multipathd reconfigure**を試すか、またはシステムを再起動してください。

18.7.4 initramfsの同期状態の維持



重要

初期RAMファイルシステム(initramfs)とブートシステムが、すべてのブロックデバイスのマルチパス処理の使用に関して一貫した動作をしていることを確認します。マルチパス設定の変更を適用した後にinitramfsを再構築します。

システムでマルチパス処理が有効になっている場合はinitramfsでも有効にする必要があり、その逆も同様です。このルールの唯一の例外は[18.3.2.2項 「ローカルディスクのルートファイルシステム」](#)のオプション**initramfs**でマルチパスを無効にします。

マルチパス設定は、ブートシステムとinitramfsの間で同期する必要があります。したがって、`/etc/multipath.conf`、`/etc/multipath/wwids`、`/etc/multipath/bindings`のいずれかのファイル、その他の設定ファイル、またはデバイスの識別に関するudevルールを変更した場合、次のコマンドを使用してinitramfsを再構築します。

```
> sudo dracut -f
```

`initramfs`とシステムが同期されていない場合、システムは正しくブートせず、起動手順を実行すると緊急シェルが起動する可能性があります。このようなシナリオを回避または修復する方法については、[18.15項 「MPIOのトラブルシューティング」](#)を参照してください。

18.7.4.1 `initramfs`でのマルチパス処理の有効化または無効化

`initramfs`が非標準的な状況で再構築される場合(カーネルパラメータの`multipath=off`を使用してブートした後やレスキューシステムからなど)、特別な注意が必要です。`dracut`は、`initramfs`の構築中に、ルートファイルシステムがマルチパスデバイス上にあることを検出した場合にのみ`initramfs`にマルチパス処理サポートを自動的に組み込みます。このような場合、マルチパス処理を明示的に有効または無効にする必要があります。

`initramfs`でマルチパスのサポートを有効にするには、次のコマンドを実行します。

```
> sudo dracut --force --add multipath
```

`initramfs`でマルチパスのサポートを無効にするには、次のコマンドを実行します。

```
> sudo dracut --force --omit multipath
```

18.7.4.2 `initramfs`の永続的なデバイス名

`dracut`が`initramfs`を生成する際には、システムが確実に正しくブートするように、永続的な方法でマウントされるディスクとパーティションを参照する必要があります。`dracut`でマルチパスデバイスを検出すると、この目的で、次のようなDM-MPデバイス名が使用されます。

```
/dev/mapper/3600a098000aad73f00000a3f5a275dc8-part1
```

これがデフォルトの動作です。これは、システムが「常に」マルチパスモードで実行されている場合には適切です。ただし、[18.7.4.1項 「`initramfs`でのマルチパス処理の有効化または無効化」](#)で説明しているようにマルチパス処理なしでシステムを起動した場合、`/dev/mapper`デバイスが存在しないため、このような`initramfs`による起動は失敗します。別の起こりうる問題シナリオと背景情報については、[18.12.4項 「マルチパスマップの参照」](#)を参照してください。

この問題が起こらないようにするには、`--persistent-policy`オプションを使用して、`dracut`の永続的なデバイス命名ポリシーを変更します。`by-uuid`使用ポリシーを設定することをお勧めします。

```
> sudo dracut --force --omit multipath --persistent-policy=by-uuid
```

手順18.1「インストール後にルートディスクのマルチパス処理を無効にする」および[18.15.2項 「デバイス参照の問題の理解」](#)も参照してください。

18.8 マルチパス設定

組み込みの `multipath-tools` は、ほとんどのセットアップでデフォルトで正しく動作します。カスタマイズが必要な場合、設定ファイルを作成する必要があります。主要設定ファイルは `/etc/multipath.conf` です。また、`/etc/multipath/conf.d/` のファイルを考慮に入れます。詳細については、[18.8.2.1項 「追加の設定ファイルと優先ルール」](#) を参照してください。

! 重要: ベンダの推奨値と組み込みハードウェアのデフォルト

一部のストレージベンダは、マルチパスオプションの推奨値をマニュアルで公開しています。これらの値は多くの場合、ベンダが自社の環境内でテストし、ストレージ製品に最適であると判断した値を表しています。[18.2.2項 「マルチパス処理のストレージアレイ自動検出」](#) の免責事項を参照してください。

`multipath-tools` には、公開されたベンダ推奨値から抽出した多くのストレージアレイ用の組み込みのデフォルトがあります。`multipath -T` を実行して、ご使用のデバイスの現在の設定を確認し、これをベンダの推奨値と比較します。

18.8.1 /etc/multipath.conf の作成

変更する設定のみが含まれている最低限の `/etc/multipath.conf` を作成することをお勧めします。多くの場合、`/etc/multipath.conf` を作成する必要はありません。

考え得る設定ディレクティブすべてを含む設定テンプレートで作業したい場合、次のコマンドを実行します。

```
multipath -T >/etc/multipath.conf
```

[18.14.1項 「設定のベストプラクティス」](#) も参照してください。

18.8.2 multipath.conf 構文

`/etc/multipath.conf` ファイルでは、セクション、サブセクション、およびオプション/値のペアの階層を使用します。

- トーカンは空白で区切られます。連続する空白文字は、引用符で囲まれていない限り1つの空白に圧縮されます(以下を参照)。
- ハッシュ(#)文字と感嘆符(!)文字により、行の残りの部分がコメントとして破棄されます。

- セクションとサブセクションは、同一行のセクション名と開き中括弧({})で開始され、独自の行の閉じ中括弧(})で終了します。
- オプションと値は1行に書き込まれます。行の継続はサポートされていません。
- オプションとセクション名はキーワードである必要があります。使用できるキーワードについては、[multipath.conf\(5\)](#)を参照してください。
- 値は二重引用符(")で囲むことができます。値に空白またはコメント文字が含まれている場合、その値を引用符で囲む必要があります。値の内側にある二重引用符文字は二重引用符のペア("")で表されます。
- 一部のオプションの値はPOSIXの正規表現です([regex\(7\)](#)を参照)。大文字と小文字が区別され、アンカーがないので「bar」は「rhabarber」と一致しますが、「Barbie」とは一致しません。

構文は、次のようにします。

```
section {
    subsection {
        option1 value
        option2      "complex value!"
        option3      "value with ""quoted"" word"
    } ! subsection end
} # section end
```

18.8.2.1 追加の設定ファイルと優先ルール

/etc/multipath.confの後、ツールは、パターン/etc/multipath.conf.d/*.confと一致するファイルを読み込みます。追加のファイルは/etc/multipath.confと同じ構文ルールに従います。セクションとオプションは複数回使用できます。「同じセクションの同じオプション」が複数のファイルで設定されたり、同じファイルの複数の行で設定されると、最後の値が優先されます。multipath.confセクション間には個別の優先ルールが適用されます。以下を参照してください。

18.8.3 multipath.confセクション

/etc/multipath.confファイルは、以下のセクションで構成されています。一部のオプションは複数のセクションで使用できます。詳細については[multipath.conf\(5\)](#)を参照してください。

defaults

一般的なデフォルト設定。



重要: 組み込みのデバイスプロパティの上書き

組み込みのハードウェア固有のデバイスプロパティはdefaultsセクションの設定より優先されます。したがって、変更は、devicesセクションまたはoverridesセクションで行う必要があります。

blacklist

無視するデバイスをリストします。[18.11.1項 「multipath.confのblacklistセクション」](#) を参照してください。

blacklist_exceptions

マルチパス処理されるデバイスをリストします(ブラックリストに含まれている場合もリストされます)。[18.11.1項 「multipath.confのblacklistセクション」](#) を参照してください。

devices

ストレージコントローラ専用の設定。このセクションはdeviceサブセクションのコレクションです。このセクションの値は、defaultsセクションと同じオプションの値、およびmultipath-toolsの組み込み設定を上書きします。

devicesセクションのdeviceエントリは、正規表現を使用して、デバイスのベンダおよび製品と照合されます。これらのエントリは「マージ」され、デバイスに対して照合するセクションのすべてのオプションが設定されます。複数の照合するdeviceセクションで同じオプションが設定されている場合、最後のデバイスエントリが優先されます。それが前のエントリよりも「特定的」でない場合でも同様です。これは、照合するエントリが異なる設定ファイルに存在する場合でも適用されます([18.8.2.1項 「追加の設定ファイルと優先ルール」](#) を参照)。次の例では、デバイスSOMEcorp STORAGEはfast_io_fail_tmo 15を使用します。

```
devices {
    device {
        vendor SOMEcorp
        product STOR
        fast_io_fail_tmo 10
    }
    device {
        vendor SOMEcorp
        product .*
        fast_io_fail_tmo 15
    }
}
```

```
    }  
}
```

multipaths

個々のマルチパスデバイスの設定。このセクションは multipath サブセクションのリストです。値は defaults セクションと devices セクションを上書きします。

上書き

他のすべてのセクションの値を上書きする設定。

18.8.4 multipath.confの変更の適用

設定変更を適用するには、次のコマンドを実行します。

```
> sudo multipathd reconfigure
```

忘れずに initramfs の設定と同期してください。18.7.4 項 「initramfs の同期状態の維持」 を参照してください。



警告: multipathを使用して設定を適用しない

multipath を実行中に multipathd コマンドを使用して新しい設定を適用しないでください。適用すると、セットアップに一貫性がなくなり、セットアップが破損する可能性があります。



注記: 変更したセットアップの確認

変更した設定を適用する前にテストできます。そのためには次のコマンドを実行します。

```
> sudo multipath -d -v2
```

このコマンドは、提案されたトポロジを使用して作成される新しいマップを表示しますが、マップが削除/フラッシュされるかどうかは表示しません。より多くの情報を得るには、詳細度を上げて次のコマンドを実行します。

```
> sudo multipath -d -v3 2>&1 | less
```

18.9 フェールオーバー、キュー格納、およびフェールバックのポリシーの設定

マルチパスI/Oの最終目標は、ストレージシステムとサーバ間のコネクティビティ耐障害性を提供することです。望ましいデフォルトの動作は、サーバがスタンドアロンのサーバか、高可用性クラスタ内のノードかによって異なります。

このセクションでは、耐障害性を実現するために最も重要なmultipath-tools設定パラメータについて説明します。

polling_interval

パスデバイスの正常性チェック間の間隔(秒単位)。デフォルトは5秒です。障害が発生したデバイスはこの間隔でチェックされます。デバイスが正常な場合、この間隔を最長max_polling_interval秒まで長くすることができます。

detect_checker

これがyes(デフォルト、推奨)に設定されている場合、multipathdは、最適なパスチェックアルゴリズムを自動的に検出します。

path_checker

パスの状態のチェックに使用するアルゴリズム。チェックカーを有効にする必要がある場合、次のようにdetect_checkerを無効にします。

```
defaults {
    detect_checker no
}
```

次のリストには、最も重要なアルゴリズムのみが含まれます。完全なリストについては、multipath.conf(5)を参照してください。

tur

TEST UNIT READYコマンドを送信します。これは、ALUAをサポートしているSCSIデバイスのデフォルトです。

directio

非同期I/O(aio)を使用してデバイスセクタを読み取ります。

rdac

NetAPP E-Seriesおよび同様のアレイ用のデバイス固有のチェックカー。

none

パスチェックは実行されません。

checker_timeout

デバイスがパスチェックコマンドに一定時間応答しない場合、失敗とみなされます。デフォルトは、デバイスに対するカーネルのSCSIコマンドのタイムアウトです(通常30秒)。

fast_io_fail_tmo

SCSIトランsportトレイヤのエラーが(たとえば、ファイバチャネルのリモートポートで)検出されると、カーネルトランsportトレイヤは、トランsportが回復するまでこの時間(秒単位)待機します。この時間経過後、パスデバイスは、「トランsportオフライン」の状態で失敗します。これは、マルチパスでは非常に有用です。マルチパスでは、頻繁に発生するクラスのエラーに対してクイックパスフェールオーバーが許可されているためです。この値は、Fabricの再設定用に一般的なタイムスケールと一致している必要があります。デフォルト値の5秒は、ファイバチャネルでは正しく動作します。iSCSIなどのその他のトランsportでは、より長いタイムアウトが必要な場合があります。

dev_loss_tmo

SCSIトランsportエンドポイント(たとえば、ファイバチャネルのリモートポート)に到達できなくなった場合、カーネルは、SCSIデバイスノードを完全に削除してポートが再表示されるまでこの時間(秒単位)待機します。デバイスノードの削除は複雑な操作であり、競合状態やデッドロックになりやすいため、避けるのが最善です。したがって、この値を大きい値に設定することをお勧めします。特別な値`infinity`がサポートされています。デフォルトは10分です。デッドロック状態を回避するために、`multipathd`は、I/Oのキュー格納(`no_path_retry`を参照)を`dev_loss_tmo`の期限が切れる前に停止します。

no_path_retry

指定マルチパスマップのすべてのパスで障害が発生した場合における処理を決定します。次の値を使用できます。

fail

マルチパスマップのI/Oは失敗します。そのため、マウントされたファイルシステムなどの上位レイヤでI/Oエラーが発生します。影響を受けるファイルシステム、および場合によってはホスト全体がデイグレードモードになります。

queue

マルチパスマップ上のI/Oがデバイスマッパーレイヤのキューに入り、パスデバイスが再び利用可能になるとそのデバイスに送信されます。これは、データ損失を回避するための最も安全なオプションですが、パスデバイスが長時間復帰しないと悪

影響を被る可能性があります。デバイスからの読み取りプロセスは中断できないスリープ(D)状態でハングします。キューに格納されたデータでメモリが一杯になり、処理できなくなります。最終的にメモリが枯渇します。

N

Nは正の整数です。N秒のポーリング間隔でマップデバイスをキュー モードのままにします。この時間が経過すると、マップデバイスのmultipathdは失敗します。polling_intervalが5秒でno_path_retryが6の場合、multipathdはI/Oをキューに約30秒間(=6 * 5秒)格納し、時間が経過するとそのマップデバイスのI/Oは失敗します。

flush_on_last_del

yesに設定し、マップのすべてのパスデバイスが(単に失敗するだけではなく)削除される場合、マップのすべてのI/Oが失敗してから、マップを削除します。デフォルトはnoです。

deferred_remove

yesに設定し、マップのすべてのパスデバイスが削除される場合、ホルダがマップデバイスのファイル記述子を閉じるのを待機してから、マップデバイスをフラッシュして削除します。最後のホルダがマップを閉じる前にパスが再表示された場合、延期された削除操作はキャンセルされます。デフォルトはnoです。

fallback

非アクティブパスグループの失敗したパスデバイスが回復すると、multipathdは、すべてのパスグループのパスグループ優先度を再評価します(18.10項「パスのグループ化と優先度の設定」を参照)。再評価後、優先度の最も高いパスグループが、現在非アクティブのパスグループの1つになる可能性があります。このパラメータによってこの状況での動作が決まります。



重要: ベンダの推奨事項に従ってください

最適なフェールバックポリシーは、ストレージデバイスの特性によって異なります。したがって、fallback設定をストレージベンダに確認することを強くお勧めします。

manual

管理者がmultipathd switchgroupを実行しない限り何も起こりません(18.6.2項「multipathd デーモン」を参照)。

immediate

優先度の最も高いパスグループがすぐにアクティブになります。これは多くの場合(特にスタンドアロンサーバでは)パフォーマンスの観点で有利ですが、パスグループの変更が負荷の大きい操作となるアレイには使用しないでください。

followover

immediateと同様ですが、アクティブになったばかりのパスのみがパスグループ内の正常なパスである場合のみフェールバックを実行します。これは、クラスタ構成の場合に有用です。別のノードが以前にフェールオーバーを要求していたときに、ノードが自動的にフェールバックしないようにします。

N

Nは正の整数です。優先度の最も高いパスグループをアクティブにする前にN個のポーリング間隔待機します。この間に優先度が再度変更されると、待機期間が新たに始まります。

eh_deadline

デバイスからの応答がなく、SCSIコマンドがエラー応答なしにタイムアウトした場合に、SCSIエラー処理に費やされる時間(秒単位)のおおよその上限を設定します。期限が経過すると、カーネルはHBAを完全にリセットします。

/etc/multipath.confファイルを変更した後、[18.8.4項 「multipath.confの変更の適用」](#)の説明に従って設定を適用します。

18.9.1 スタンドアロンサーバでのキュー policy

スタンドアロンサーバに対してマルチパスI/Oを設定する際は、no_path_retryで値をqueueに設定することにより、サーバのオペレーティングシステムを、I/Oエラーの受信から可能な限り保護することができます。この設定では、マルチパスのフェールオーバーが発生するまでメッセージはキューに入ります。「無限」のキューを望まない場合(上記を参照)、通常の状況下でストレージパスが回復するのに十分大きいとみなされる数値を選択します(上記を参照)。

18.9.2 クラスタ化されたサーバでのキュー policy

高可用性クラスタ内のノードに対してマルチパスI/Oを構成するときには、マルチパスドリソースのフェールオーバーをトリガするためにI/O障害が報告されるようにして、マルチパスのフェールオーバーが解決されるのを待たなくて済むようにするとよいでしょう。クラスタ環境では、no_path_retry 設定を、ストレージシステムへの接続が失われた場合に、クラスタ

ノードがクラスタ検証プロセスに関連するI/Oエラー(ハートビート許容値の50%を推奨)を受信するように変更する必要があります。また、パスの障害によるリソースのピンポンを避けるため、マルチパスの`fallback`を`manual`または`followover`に設定するとよいでしょう。

18.10 パスのグループ化と優先度の設定

マルチパスマップのパスデバイスは、「パスグループ」(「優先度グループ」とも呼ばれる)にグループ化されます。常に1つのパスグループのみがI/Oを受信します。`multipathd`は、「優先度」をパスグループに割り当てます。アクティブなパスを持つパスグループの内、優先度の最も高いグループが、マップに設定されたフェールバックポリシーに応じてアクティブになります([18.9項 「フェールオーバー、キュー格納、およびフェールバックのポリシーの設定](#)」を参照)。パスグループの優先度は、パスグループ内のアクティブなパスデバイスの優先度の平均です。パスデバイスの優先度は、デバイスのプロパティから計算される整数値です(以下の`prio`オプションの説明を参照)。

このセクションでは、優先度の決定とパスのグループ化に関する`multipath.conf`設定パラメータについて説明します。

path_grouping_policy

パスをグループに結合するために使用する方法を指定します。最も重要なポリシーのみがここにリストされます。使用頻度の低いその他の値については、`multipath.conf(5)`を参照してください。

failover

パスグループごとに1つのパス。この設定は、従来の「アクティブ/パッシブ」ストレージアレイで有用です。

multibus

1つのパスグループ内のすべてのパス。これは、従来の「アクティブ/パッシブ」アレイで有用です。

group_by_prio

同じパス優先度のパスデバイスがグループ化されます。このオプションは、ALUAのように非対称アクセス状態をサポートする最新のアレイで有用です。`alua`または`sysfs`の優先度アルゴリズムと組み合わせて、`multipathd`によってセットアップされる優先度グループは、ストレージアレイがALUA関連のSCSIコマンドを使用してレポートするプライマリターゲットポートグループと照合されます。

マルチパスマップのパスグループ化ポリシーは、同じポリシー名を使用して、次のコマンドで一時的に変更できます。

```
> sudo multipath -p POLICY_NAME MAP_NAME
```

marginal_pathgroups

onまたはfpinに設定すると、「ぎりぎり」のパスデバイスは、別のパスグループに分類されます。これは、使用しているパスグループ化アルゴリズムとは無関係です。[18.13.1 項「信頼性の低い\(「ぎりぎりの」\)パスデバイスの処理」](#)を参照してください。

detect_prio

これがyes (デフォルト、推奨)に設定されている場合、multipathdはストレージデバイスの優先度を設定するのに最適なアルゴリズムを自動的に検出して、prio設定を無視します。実際には、これは、ALUAのサポートが検出された場合にsysfs prioアルゴリズムを使用することを意味します。

prio

パスデバイスの優先度を導出する方法を決定します。これを上書きする場合、次のようにdetect_prioを無効にします。

```
defaults {
    detect_prio no
}
```

次のリストには、最も重要な方法のみが含まれます。他にもいくつかの方法が利用可能です。これらは主にレガシハードウェアをサポートするためのものです。完全なリストについては、[multipath.conf\(5\)](#)を参照してください。

alua

SCSI-3 ALUAアクセス状態を使用して、パス優先値を導出します。オプションのexclusive_pref_bit引数を使用して、ALUAの「優先プライマリターゲットポートグループ」(PREF)ビットが設定されているデバイスの動作を変更できます。

```
prio alua
prio_args exclusive_pref_bit
```

このオプションを設定すると、「優先」パスは、他のアクティブ/最適化パスを上回る優先度ボーナスを獲得します。このオプションを設定しないと、すべてのアクティブ/最適化パスは、同じ優先度を割り当てられます。

sysfs

aluaと似ていますが、SCSIコマンドをデバイスに送信する代わりに、sysfsからアクセス状態を取得します。これにより、aluaよりもI/O負荷が軽減されますが、ALUAがサポートしているすべてのストレージアレイに適しているわけではありません。

const

すべてのパスに定数値を使用します。

path_latency

パスデバイスでI/Oレイテンシ(I/O送信から完了までの時間)を測定し、低レイテンシのデバイスに高い優先度を割り当てます。詳細については[multipath.conf\(5\)](#)を参照してください。このアルゴリズムはまだ実験段階です。

weightedpath

名前、シリアル番号、Host:Bus:Target:Lun ID (HBTL)、またはファイバチャネルWWNに基づいて優先度をパスに割り当てます。優先度の値は時間が経過しても変化しません。この方法では、引数が必要です。詳細については、[multipath.conf\(5\)](#)を参照してください。例:

```
prio weightedpath  
prio_args "hbtl 2:.*:.*:.* 10 hbtl 3:.*:.*:.* 20 hbtl .* 1"
```

これは、SCSIホスト3のデバイスにSCSIホスト2のデバイスより高い優先度を割り当てる、他のすべてに低い優先度を割り当てています。

prio_args

一部のアルゴリズムは、追加の引数が必要です。これらはこのオプションで指定されます。構文はアルゴリズムによって決まります。詳細については上記の説明を参照してください。

hardware_handler

パスグループを切り替えるときにパスデバイスをアクティブ化するためにカーネルが使用するカーネルモジュールの名前。このオプションは、最新のカーネルでは効果がありません。最新のカーネルでは、ハードウェアハンドラが自動検出されるためです。[18.2.3項 「特定のハードウェアハンドラを必要とするストレージアレイ」](#)を参照してください。

path_selector

アクティブパスグループのパス間のロードバランシングに使用するカーネルモジュールの名前。利用可能な選択肢は、カーネル設定によって決まります。過去の経緯から、[multipath.conf](#)では名前は常に引用符で囲み、その後に「0」を付ける必要があります。次に例を示します。

```
path_selector "queue-length 0"
```

service-time

すべてのパスで保留中のI/Oが完了するために必要な時間を推定し、最小値のパスを選択します。これがデフォルトの設定です。

historical-service-time

過去のサービス時間(移動平均を保持する時間)、および未処理の要求数に基づいて、将来のサービス時間を推定します。すべてのパスで保留中のI/Oが完了するために必要な時間を推定し、最小値のパスを選択します。

queue-length

現在保留中のI/O要求が最も少ないパスを選択します。

round-robin

ラウンドロビン方式でパスを切り替えます。次のパスに切り替わる前に、そのパスに送信される要求数は、rr_min_io_rqとrr_weightオプションで調整できます。

io-affinity

このパスセレクタは、現時点ではmultipath-toolsでは機能しません。

/etc/multipath.confファイルを変更した後、18.8.4項「multipath.confの変更の適用」の説明に従って設定を適用します。

18.11 マルチパス処理のためのデバイスの選択

マルチパスデバイスを含むシステムでは、一部のデバイス(通常はローカルディスク)でのマルチパスマップのセットアップを避けた方がよい場合があります。multipath-toolsは、どのデバイスをマルチパスのパスデバイスとみなすかを設定するためのさまざまな方法を提供します。



注記: ローカルディスク上のマルチパス

通常、ローカルディスクの上にデバイスが1つだけある「デイグレード」マルチパスマップをセットアップしても問題ありません。これは正常に動作し、追加の設定は不要です。しかし、管理者の中にはこれを分かりにくく感じたり、一般的にこの種の不必要的マルチパス処理に反対したりする人もいます。また、マルチパスレイヤを使用すると、パフォーマンスにわずかなオーバーヘッドが生じます。18.3.2.2項「ローカルディスクのルートファイルシステム」も参照してください。

/etc/multipath.confファイルを変更した後、18.8.4項「multipath.confの変更の適用」の説明に従って設定を適用します。

18.11.1 multipath.confのblacklistセクション

/etc/multipath.confファイルには、multipathdとmultipathで無視する必要があるすべてのデバイスを列挙するblacklistセクションを含めることができます。次の例は、デバイスを除外する考え方を示しています。

```
blacklist {
    wwid 3600605b009e7ed501f0e45370aaeb77f ❶
    device { ❷
        vendor ATA
        product .*
    }
    protocol scsi:sas ❸
    property SCSI_IDENT_LUN_T10 ❹
    devnode "!^dasd[a-z]*" ❺
}
```

- ❶ wwidエントリは、ルートディスクなどの特定のデバイスを除外するのに最適です。
- ❷ このdeviceセクションは、すべてのATAデバイスを除外します(productの正規表現はすべてに一致します)。
- ❸ protocolによる除外では、特定のバスタイプ(ここではSAS)を使用してデバイスを除外できます。他の一般的なプロトコル値はscsi:fcp、scsi:iscsi、およびccwです。詳細については、multipath.conf(5)を参照してください。システムのパスが使用しているプロトコルを表示するには、次のコマンドを実行します。

```
> sudo multipathd show paths format "%d %P"
```

この形式はSLES 15 SP1およびSLES 12 SP5以降でサポートされています。

- ❹ このpropertyエントリは、特定のudevプロパティを含むデバイスを除外します(プロパティの値は無関係)。
- ❺ devnodeによるデバイスの除外は、例のように正規表現を使用するデバイスのクラスに対してのみ推奨されます。この例では、DASDデバイス以外のすべてを除外します。sdaのような個々のデバイスにこれを使用することは、デバイスのノード名が永続的ではないため、お勧めしません。

この例ではblacklistおよびblacklist_exceptionsセクションでのみサポートされる特殊な構文を示しています。正規表現の前に感嘆符(!)を付けると、一致が否定されます。

感嘆符は二重引用符内に配置される必要があることに注意してください。

デフォルトでは、multipath-toolsは、SCSI、DASDまたはNVMeを除くすべてのデバイスを無視します。技術的には、組み込みのdevnode除外リストは、次の否定された正規表現です。

```
devnode !^(sd[a-z]|dasd[a-z]|nvme[0-9])
```

18.11.2 `multipath.conf` の `blacklist exceptions` セクション

特定のデバイスのみをマルチパス処理用に設定することが望ましい場合があります。この場合、デバイスはデフォルトで除外され、マルチパスマップの一部となるデバイスに対して例外が定義されます。この目的で `blacklist_exceptions` セクションが存在します。これは通常、次の例のように使用されます。この例では、製品文字列「NETAPP」のストレージを除くすべてが除外されます。

```
blacklist {
    wwid .
}
blacklist_exceptions {
    device {
        vendor ^NETAPP$
        product .
    }
}
```

`blacklist_exceptions` セクションでは、上記の `blacklist` セクションで説明されているすべての方法がサポートされます。

`blacklist_exceptions` の `property` ディレクティブは必須です。なぜなら、マルチパスのパスデバイスとみなされるために、すべてのデバイスは少なくとも1つの「許可された」udevプロパティを持つ「必要がある」ためです(プロパティの値は重要ではありません)。`property` の組み込みのデフォルトは次のとおりです。

```
property (SCSI_IDENT_|ID_WWN)
```

この正規表現に一致するudevプロパティを少なくとも1つ持つデバイスのみが含まれます。

18.11.3 デバイスの選択に影響するその他のオプション

`blacklist` オプション以外にも、`/etc/multipath.conf` には、どのデバイスをマルチパスのパスデバイスとみなすかに影響するその他の設定がいくつかあります。

`find_multipaths`

このオプションは、除外されないデバイスが最初に検出された場合の `multipath` と `multipathd` の動作を制御します。次の値を使用できます。

greedy

`/etc/multipath.conf`の`blacklist`によって除外されないすべてのデバイスが含まれます。これは、SUSE Linux Enterpriseのデフォルトです。この設定がアクティブの場合、マルチパスマップへのデバイスの追加を防ぐ唯一の方法は、デバイスを除外と設定することです。

strict

すべてのデバイスは、`/etc/multipath.conf`の`blacklist`セクションに存在しなくても、そのWWIDが`/etc/multipath/wwids`ファイルにリストされていない限り、除外されます。これには、WWIDファイルの手動メンテナンスが必要です(下記の注意事項を参照)。

yes

デバイスは、`strict`の条件を満たしている場合、または同じWWIDを持つ他のデバイスがシステム内に少なくとも1つ存在する場合に含まれます。

smart

新しいWWIDを初めて検出した場合、これは、マルチパスのパスデバイスとして一時的にマークされます。`multipathd`は、同じWWIDを持つ追加のパスが現れるまでしばらく待機します。現れると、マルチパスマップは通常どおりにセットアップされます。それ以外の場合、タイムアウトになると、1つのデバイスが非マルチパスデバイスとしてシステムにリリースされます。タイムアウトは、オプション`find_multipaths_timeout`を使用して設定できます。

このオプションは、SUSE Linux Enterprise Server 15でのみ利用できる`systemd`機能に依存します。



注記: /etc/multipath/wwidsの管理

`multipath-tools`は、前にセットアップしたマルチパスマップの記録をファイル`/etc/multipath/wwids`(「WWIDsファイル」)に保持します。このファイルにリストされているWWIDを持つデバイスは、マルチパスのパスデバイスとみなされます。このファイルは、`greedy`を除く`find_multipaths`のすべての値に対して、マルチパスデバイスの選択に重要です。

`find_multipaths`が`yes`または`smart`に設定されている場合、`multipathd`は、新しいマップをセットアップした後にWWIDを`/etc/multipath/wwids`に追加します。これにより、今後これらのマップはより迅速に検出されるようになります。

WWIDファイルは、手動で変更できます。

```
> sudo multipath -a 3600a098000aad1e3000064e45f2c2355 ①
```

```
> sudo multipath -w /dev/sdf ②
```

- ① このコマンドを実行すると、指定されたWWIDが`/etc/multipath/wwids`に追加されます。
- ② このコマンドを実行すると、指定されたデバイスのWWIDが削除されます。
`strict`モードでは、これが新しいマルチパスデバイスを追加する唯一の方法です。WWIDファイルを変更した後、`multipathd reconfigure`を実行して変更を適用します。変更をWWIDファイルに適用した後にinitramfsを再構築することをお勧めします([18.7.4項 「initramfsの同期状態の維持」](#)を参照)。

allow_usb_devices

このオプションが`yes`に設定されている場合、USBストレージデバイスはマルチパス処理用とみなされます。デフォルトは`no`です。

18.12 マルチパスデバイス名とWWID

`multipathd`と`multipath`は、WWIDを内部的に使用して、デバイスを識別します。WWIDは、デフォルトではマップ名としても使用されます。便宜上、`multipath-tools`は、よりシンプルで簡単に覚えることができる名前のマルチパスデバイスへの割り当てをサポートしています。

18.12.1 WWIDとデバイスの識別

マルチパス操作では、同じストレージボリュームへのパスを表すデバイス検出の信頼性が高いことが非常に重要です。`multipath-tools`は、この目的でデバイスのWWID (World Wide Identification: ワールドワイドID)を使用します(これはUUID (Universally Unique ID: ユニバーサルユニークIDまたはUID (Unique ID: ユニークID)と呼ばれることもあります(ユーザIDと混同しないでください))。マップデバイスのWWIDは、常にそのパスデバイスのWWIDと同じです。

デフォルトでは、パスデバイスのWWIDは、`sysfs`ファイルシステムからデバイスの属性を読み取ることによって、または固有のI/Oコマンドを使用して、`udev`ルールで決定されるデバイスの`udev`プロパティから推測されます。デバイスの`udev`プロパティを確認するには、次のコマンドを実行します。

```
> udevadm info /dev/sdx
```

WWIDを導出するために`multipath-tools`によって使用される`udev`プロパティを次に示します。

- SCSIデバイスのID_SERIAL(これをデバイスの「シリアル番号」と混同しないでください)
- DASDデバイスのID_UID
- NVMeデバイスのID_WWN



警告: WWIDの変更を回避する

使用中のマルチパスマップのWWIDは変更できません。設定の変更により、マップされたパステーブルのWWIDが変更された場合は、マップを破棄し、新しいWWIDで新しいマップをセットアップする必要があります。これは、古いマップを使用中には実行できません。極端な場合、WWIDの変更によってデータが破損する可能性があります。したがって、マップのWWIDが変更される結果となる設定変更の適用は、「厳格に回避する」必要があります。

`/etc/multipath.conf`でuid_attrsオプションを有効にすることは許可されています。[18.13項 「その他のオプション」](#)を参照してください。

18.12.2 マルチパスマップのエイリアスの設定

`/etc/multipath.conf`のmultipathsセクションで、次のように任意のマップ名を設定できます。

```
multipaths {
    multipath {
        wwid 3600a098000aad1e3000064e45f2c2355
        alias postgres
    }
}
```

エイリアスは表現力が豊かですが、各マップに個別に割り当てる必要があるため、大規模なシステムでは面倒な場合があります。

18.12.3 自動生成されるユーザフレンドリ名の使用

`multipath-tools`は、自動生成されたエイリアス、いわゆる「ユーザフレンドリ名」もサポートしています。エイリアスの命名規則は、`mpathINDEX`というパターンに従います。ここで_{INDEX}は小文字です(aで始まります)。したがって、最初に自動生成されるエイリアスは`mpatha`で、その後`mpathb`、`mpathc`と続き、`mpathz`まで続きます。その後は、`mpathaa`、`mpathab`などが続けます。

マップ名は、永続的である場合のみ有用です。`multipath-tools`は、`/etc/multipath/bindings`ファイル(「バインディングファイル」)で割り当てられた名前を追跡します。新しいマップが作成されると、最初にこのファイルでWWIDが検索されます。WWIDが見つからない場合、最も可用性が低いユーザフレンドリ名がこれに割り当てられます。

18.12.2項 「マルチパスマップのエイリアスの設定」で説明されているように、明示的なエイリアスがユーザフレンドリ名よりも優先されます。

`/etc/multipath.conf`の次のオプションは、ユーザフレンドリ名に影響します。

user_friendly_names

`yes`に設定すると、ユーザフレンドリ名が割り当てられ、使用されます。それ以外の場合、エイリアスが設定されていない限り、WWIDがマップ名として使用されます。

alias_prefix

ユーザフレンドリ名を作成するために使用するプレフィクス。デフォルトでは`mpath`です。



警告: 高可用性クラスタのマップ名

クラスタ操作では、クラスタ内のすべてのノードにわたってデバイス名と同じにする必要があります。`multipath-tools`の設定は、ノード間で同期状態を保つ必要があります。`user_friendly_names`を使用している場合、`multipathd`は実行時に`/etc/multipath/bindings`ファイルを変更する可能性があります。このような変更は、すべてのノードに対して動的に複製する必要があります。同じことが`/etc/multipath/wwids`にも当てはまります(18.11.3項 「デバイスの選択に影響するその他のオプション」を参照)。



注記: 実行時のマップ名の変更

実行時にマップ名を変更できます。このセクションで説明されている方法のいずれかを使用して、`multipathd reconfigure`を実行すると、システム操作を妨げることなくマップ名が変更されます。

18.12.4 マルチパスマップの参照

技術的には、マルチパスマップは、デバイスマッパーデバイスです。これには、`/dev/dm-n`という形式の汎用名があります(`n`は整数)。これらの名前は永続的ではありません。これらは、マルチパスマップを参照するために使用「しない」でください。`udev`を実行すると、これら

のデバイスへのさまざまなシンボリックリンクが作成されます。これらは、永続的な参照としてより適切です。これらのリンクは、特定の設定変更に対する不变性の点で異なります。次の一般的な例は、すべて同じデバイスを指しているさまざまなシンボリックリンクを示しています。

```
/dev/disk/by-id/dm-name-mpathb① -> ../../dm-1  
/dev/disk/by-id/dm-uuid-mpath-3600a098000aad73f00000a3f5a275dc8② -> ../../dm-1  
/dev/disk/by-id/scsi-3600a098000aad73f00000a3f5a275dc8③ -> ../../dm-1  
/dev/disk/by-id/wwn-0x600a098000aad73f00000a3f5a275dc8④ -> ../../dm-1  
/dev/mapper/mpathb⑤ -> ../../dm-1
```

- ① ⑤これら2つのリンクは、マップ名を使用してマップを参照します。したがって、マップ名が変更されるとリンクも変更されます(たとえば、ユーザフレンドリ名を有効または無効にした場合)。
- ② このリンクは、デバイスマッパーUUIDを使用します。これは、[multipath-tools](#)によって使用されているWWIDにプレフィックスとして文字列dm-uuid-mpath-が付いています。これはマップ名とは無関係です。
デバイスマッパーUUIDは、確実に「マルチパスデバイスのみ」が参照されるようにするために推奨される形式です。たとえば、/etc/lvm/lvm.confの次の行では、マルチパスマップを除くすべてのデバイスが拒否されます。

```
filter = [ "a|/dev/disk/by-id/dm-uuid-mpath-.*|", "r|.*|" ]
```

- ③ ④これらは、通常パスデバイスを指すリンクです。マルチパスデバイスは、udevリンクの優先度が高いため、これらを引き継ぎました(udev(7)を参照)。マップが破棄されたり、マルチパス処理がオフになったりしても、それらは引き続き存在し、代わりにパスデバイスのいずれかを指します。これは、マルチパス処理がアクティブかどうかに関係なく、WWIDによってデバイスを参照する手段を提供します。

kpartxツールで作成されたマルチパスマップ上の「パーティション」の場合、親デバイス名またはWWIDとパーティション番号から派生した同様のシンボリックリンクがあります。

```
/dev/disk/by-id/dm-name-mpatha-part2 -> ../../dm-5  
/dev/disk/by-id/dm-uuid-part2-mpath-3600a098000aad1e300000b4b5a275d45 -> ../../dm-5  
/dev/disk/by-id/scsi-3600a098000aad1e300000b4b5a275d45-part2 -> ../../dm-5  
/dev/disk/by-id/wwn-0x600a098000aad1e300000b4b5a275d45-part2 -> ../../dm-5  
/dev/disk/by-partuuid/1c2f70e0-fb91-49f5-8260-38eacaf7992b -> ../../dm-5  
/dev/disk/by-uuid/f67c49e9-3cf2-4bb7-8991-63568cb840a4 -> ../../dm-5  
/dev/mapper/mpatha-part2 -> ../../dm-5
```

パーティションにはby-uuidリンクもあることが多く、デバイス自体を参照するのではなく、デバイスに含まれているファイルシステムを参照します。多くの場合、これらのリンクが推奨されます。これらは、ファイルシステムが別のデバイスまたはパーティションにコピーされても不变です。



警告: initramfsのマップ名

`dracut`は、initramfsを構築するとき、initramfsのデバイスへのハードコード化された参照を作成します。デフォルトでは、`/dev/mapper/$MAP_NAME`参照を使用します。これらのハードコード化された参照は、initramfsで使用されているマップ名がinitramfs構築中に使用した名前と一致しない場合、起動中には見つからず、起動エラーになります。通常、この状況にはなりません。その理由は、`dracut`がすべてのマルチパス設定ファイルをinitramfsに追加するためです。ただし、initramfsが異なる環境(レスキュースистем、またはオフライン更新中など)から構築されている場合、この問題が発生する可能性があります。この起動エラーが発生しないようにするには、[18.7.4.2項 「initramfsの永続的なデバイス名」](#)で説明されているように、`dracut`の`persistent_policy`設定を変更します。

18.13 その他のオプション

このセクションでは、これまでに説明しなかった有用ないくつかの`multipath.conf`オプションを一覧表示します。完全なリストについては、`multipath.conf(5)`を参照してください。

verbosity

`multipath`と`multipathd`両方のログ詳細度を制御します。コマンドラインオプション`-v`を使用すると、両方のコマンドについてこの設定が上書きされます。この値は、0(致命的なエラーのみ)と4(詳細なログ記録)の間で設定できます。デフォルトは2です。

uid_attrs

このオプションを使用すると、udevイベントの処理を最適化できます(いわゆる「ueventマージ」)。これは、数百のパスデバイスが同時に障害を起こしたり同時に再検出されたりする環境で有用です。パスのWWIDが変更されないようにするためにには([18.12.1項 「WWIDとデバイスの識別」](#)を参照)、この値を正確に次のように設定する必要があります。

```
defaults {
    uid_attrs "sd:ID_SERIAL dasd:ID_UID nvme:ID_WWN"
}
```

skip_kpartx

マルチパスデバイスでyesに設定する場合(デフォルトはno)、指定したデバイスの上にパーティションデバイスを作成しないでください([18.7.3項 「マルチパスデバイスのパーティションとkpartx」](#) を参照)。仮想マシンによって使用されるマルチパスデバイスで有用です。以前のSUSE Linux Enterprise Serverリリースでは、同じ効果をパラメータ「`features 1 no_partitions`」を使用して実現していました。

max_sectors_kb

マルチパスマップのすべてのパスデバイスに対して1つのI/O要求で送信される最大データ量を制限します。

ghost_delay

アクティブ/パッシブアレイで、アクティブパスの前にパッシブパス(「ゴースト」状態)が検出される場合があります。マップが直ちにアクティブ化され、I/Oが送信された場合、これにより、負荷の大きいパスのアクティブ化が実行される可能性があります。このパラメータは、マップをアクティブにする前に、マップのアクティブパスが表示されるまでに待機する時間(秒単位)を指定します。デフォルトはno(ゴースト遅延なし)です。

recheck_wwid

yesに設定されている場合(デフォルトはno)、障害発生後に復元されたパスのWWIDをダブルチェックし、WWIDが変更された場合はパスを削除します。これはデータの破損を防ぐための安全対策です。

enable_foreign

`multipath-tools`は、デバイスマッパーのマルチパス以外のマルチパス処理バックエンド用にプラグインAPIを提供します。このAPIは、`multipath -ll`のような標準コマンドを使用して、マルチパストポロジに関する情報の監視および表示をサポートしています。トポロジの変更はサポートされていません。

`enable_foreign`の値は、外部のライブラリ名と照合するための正規表現です。デフォルト値は「`NONE`」です。

SUSE Linux Enterprise Serverには、ネイティブNVMeマルチパス処理のサポートを追加する`nvme`プラグインが備わっています([18.2.1項 「マルチパスの実装: デバイスマッパーとNVMe」](#) を参照)。`nvme`プラグインを有効にするには、次のように設定します。

```
defaults {
    enable_foreign nvme
}
```

18.13.1 信頼性の低い(「ぎりぎりの」)パスデバイスの処理

Fabricの状態が不安定だと、パスデバイスの動作が不安定になる可能性があります。I/Oエラーの発生、回復、障害の再発が頻繁に起こります。このようなパスデバイスは、「ぎりぎりの」または「不安定な」パスと呼ばれることもあります。このセクションでは、この問題に対処するためにmultipath-toolsが提供するオプションをまとめています。



注記: multipathdのぎりぎりのパスチェックアルゴリズム

パスデバイスで、最初の失敗が発生してから`marginal_path_double_failed_time`が経過する前に、2回目の失敗(良好→不良の移行)が発生した場合、`multipathd`は、監視期間`marginal_path_err_sample_time`の間、1秒あたり10回の要求の速度でパスの監視を開始します。監視期間中のエラー率が`marginal_path_err_rate_threshold`を超えると、このパスはぎりぎりに分類されます。`marginal_path_err_recheck_gap_time`経過後、パスは、再び通常状態に移行します。

このアルゴリズムは、4つの数値パラメータ`marginal_path_`がすべて正の値に設定されていて、`marginal_pathgroups`が`fpin`に設定されていない場合に使用されます。これは、SUSE Linux Enterprise Server 15 SP1およびSUSE Linux Enterprise Server 12 SP5以降で利用できます。

`marginal_path_double_failed_time`

パス監視をトリガする2回のパスの失敗間の最大時間(秒単位)。

`marginal_path_err_sample_time`

パス監視間隔の長さ(秒単位)。

`marginal_path_err_rate_threshold`

最小エラー率(I/O 1000回あたり)。

`marginal_path_err_recheck_gap_time`

パスをぎりぎりの状態に保つ時間(秒単位)。

`marginal_pathgroups`

このオプションは、SLES 15SP3以降使用できます。次の値を使用できます。

off

ぎりぎりの状態は**multipathd**によって決まります(上記を参照)。ぎりぎりのパスは、ぎりぎりの状態である限り、復帰しません。これはデフォルトであり、**marginal_pathgroups**オプションが利用できなかったSUSE Linux Enterprise ServerのSP3より前のリリースにおける動作です。

on

offオプションと似ていますが、ぎりぎりのパスを失敗状態のままにするのではなく、別のパスグループに移動します。このグループは、他のすべてのパスグループより低い優先度を割り当てられます([18.10項 「パスのグループ化と優先度の設定](#) を参照)。このパスグループのパスは、他のパスグループのすべてのパスが失敗した場合のみI/Oに使用されます。

fpin

この設定は、SLES 15SP4以降使用できます。ぎりぎりのパス状態は、FPINイベントから導出されます(以下を参照)。ぎりぎりのパスは、**off**の説明と同様に別のパスグループに移動します。この設定では、ホスト側の追加設定は不要です。これは、FPINをサポートしているファイバチャネルファブリックでぎりぎりのパスを処理する場合に推奨される方法です。



注記: FPINベースのぎりぎりのパス検出

multipathdは、FPIN (Fibre Channel Performance Impact Notifications: ファイバチャネルのパフォーマンスへの影響通知)をリスンします。パスデバイスに対してFPIN-LI (リンクの完全性)イベントを受信すると、パスは、ぎりぎりの状態になります。この状態は、RSCNまたはリンクアップイベントが、デバイスの接続先のファイバチャネルアダプタで受信されるまで継続します。

パラメータ**san_path_err_threshold**、**san_path_err_forget_rate**、および**san_path_err_recovery_time**を使用したより単純なアルゴリズムも利用可能で、SUSE Linux Enterprise Server 15 (GA)で推奨されています。**multipath.conf(5)**の「不安定なパスの検出」セクションを参照してください。

18.14 ベストプラクティス

18.14.1 設定のベストプラクティス

設定ディレクティブの数が多いと最初は気後れします。通常、クラスタ環境で作業していない限り、空の設定で良い結果を得ることができます。

ここでは、スタンドアロンサーバの一般的な推奨事項を紹介します。これらは「必須ではありません」。背景情報については、前の各セクションの該当するパラメータのマニュアルを参照してください。

```
defaults {
    deferred_remove      yes
    find_multipaths     smart
    enable_foreign       nvme
    marginal_pathgroups fpin   # 15.4 only, if supported by fabric
}
devices {
    # A catch-all device entry.
    device {
        vendor          .*
        product         .*
        dev_loss_tmo    infinity
        no_path_retry   60           # 5 minutes
        path_grouping_policy group_by_prio
        path_selector    "historical-service-time 0"
        reservation_key file        # if using SCSI persistent reservations
    }
    # Follow up with specific device entries below, they will take precedence.
}
```

/etc/multipath.confファイルを変更した後、18.8.4項 「multipath.confの変更の適用」 の説明に従って設定を適用します。

18.14.2 マルチパスI/Oステータスの解釈

マルチパスサブシステムの概要を簡単に確認するには、`multipath -ll`または`multipathd show topology`を使用します。これらのコマンドの出力の形式は同じです。前のコマンドはカーネルの状態を読み取り、後のコマンドはマルチパスデーモンのステータスを出力します。通常、両方の状態は同じです。出力の一例を次に示します。

```
> sudo multipathd show topology
mpatha① (3600a098000aad1e300000b4b5a275d45②) dm-0③ NETAPP,INF-01-00④
```

```

size=64G features='3 queue_if_no_path pg_init_retries 50' ⑤ hwhandler='1 alua' ⑥ wp=rw ⑦
| +- ⑧ policy='historical-service-time 2' ⑨ prio=50 ⑩ status=active ⑪
| | - ⑫ 3:0:0:1 ⑬ sdb 8:16 ⑭ active ⑮ ready ⑯ running ⑰
| ` - 4:0:0:1 sdf 8:80 active ready running
`-- policy='historical-service-time 2' prio=10 status=enabled
   - 4:0:1:1 sdj 8:144 active ready running

```

- ① マップ名。
- ② マップのWWID(マップ名と違う場合)。
- ③ マップデバイスのデバイスノード名。
- ④ ベンダと製品名。
- ⑤ パスグループ。パスグループの下のインデントされた行には、このグループに属しているパスデバイスがリストされます。
- ⑥ パスグループで使用されるパスセレクタアルゴリズム。「2」は無視できます。
- ⑦ パスグループの優先度。
- ⑧ パスグループのステータス(active、enabledまたはdisabled)。アクティブなパスグループは、I/Oが現在送信されているグループです。
- ⑨ パスデバイス。
- ⑩ デバイスのバスID(ここでは、SCSI Host:Bus:Target:Lun ID)。
- ⑪ デバイスノード名とパスデバイスのメジャー/マイナー番号。
- ⑫ パスのカーネルデバイスマッパー状態(activeまたはfailed)。
- ⑬ マルチパスのパスデバイスの状態(以下を参照)。
- ⑭ カーネルのパスデバイスの状態。これはデバイスタイプ固有の値です。SCSIでは、これはrunningまたはofflineのいずれかです。

マルチパスのパスデバイスの状態を次に示します。

<u>ready</u>	パスは正常で、稼働しています
<u>ghost</u>	アクティブ/パッシブアレイのパッシブパス
<u>faulty</u>	パスがダウンしているか、到達できません
<u>i/o timeout</u>	チェックコマンドがタイムアウトしています
<u>i/o pending</u>	パスチェックコマンドの完了を待っています

<u>delayed</u>	「フラッピング」を回避するためにパスの再インスタンス化が遅延しています
<u>shaky</u>	信頼性の低いパス(emcパスチェックのみ)

18.14.3 マルチパスデバイスでのLVM2の使用

LVM2には、マルチパスデバイスを検出するためのサポートが組み込まれています。これは/etc/lvm/lvm.confでデフォルトでアクティブになっています。

```
multipath_component_detection=1
```

これは、デバイスのプロパティに関する情報もudevから取得するようにLVM2が設定されている場合のみ安定して動作します。

```
external_device_info_source="udev"
```

これは、SUSE Linux Enterprise 15 SP4ではデフォルトですが、以前のリリースでは違います。マルチパスデバイスを除くすべてのデバイスを無視するLVM2のフィルタ式を作成することも可能です(通常は不要です)。[18.12.4項 「マルチパスマップの参照](#)」を参照してください。

18.14.4 停止したI/Oの解決

すべてのパスで同時に障害が発生し、I/Oがキューに格納される場合、アプリケーションが長時間停止する可能性があります。これを解決するために、次の手順を使用できます。

1. 端末のプロンプトで、次のコマンドを入力します。

```
> sudo multipathd disablequeueing map MAPNAME
```

MAPNAMEをデバイスの正しいWWIDまたはマップされたエイリアス名で置き換えます。このコマンドにより、待ち行列で待機中のすべてのI/Oがエラーとなり、エラーが呼び出し側アプリケーションにプロパゲートします。ファイルシステムはI/Oエラーを監視し、読み込み専用モードに切り替わります。

2. 次のコマンドを入力して、キューを再びアクティブにします。

```
> sudo multipathd restorequeueing MAPNAME
```

18.14.5 マルチパスデバイスのMD RAID

マルチパス処理の上部でMD RAIDアレイは、システムのudevルールによって自動的にセットアップされます。/etc/mdadm.confの特別な設定は不要です。

18.14.6 新規デバイスのスキャン(再起動なし)

ご使用のシステムがマルチパス処理用に設定されており、SANにストレージを追加する必要がある場合は、rescan-scsi-bus.shスクリプトを使用して新しいデバイスをスキャンすることができます。コマンドの一般的な構文は次の例に従います。

```
> sudo rescan-scsi-bus.sh [-a] [-r] --hosts=2-3,5
```

各オプションには次のような意味があります。

-a

このオプションを使用すると、すべてのSCSIターゲットがスキャンされます。使用しないと、既存のターゲットのみが新しいLUNに対してスキャンされます。

-r

このオプションを使用すると、ストレージ側で削除されたデバイスを削除できます。

--hosts

このオプションを使用すると、スキャンするホストバスアダプタのリストが指定されます(デフォルトはすべてスキャン)。

その他のオプションのヘルプを表示するには、rescan-scsi-bus.sh --helpを実行します。

multipathdが実行中で、新しいSANデバイスが検出される場合、[18.11項 「マルチパス処理のためのデバイスの選択」](#)で説明されている設定に従って、マルチパスマップとして自動的にセットアップされるはずです。



警告: Dell/EMC PowerPath環境

EMC PowerPath環境では、SCSIバスをスキャンする場合に、オペレーティングシステムに付属するrescan-scsi-bus.shユーティリティまたはHBAベンダスクリプトを使用しないでください。ファイルシステムが破損する可能性を避けるため、EMCでは、Linux用EMC PowerPathのベンダマニュアルに記載されている手順に従うよう求めています。

18.15 MPIOのトラブルシューティング

マルチパスを含むシステムでシステムが緊急モードになり、見つからないデバイスに関するメッセージが出力される場合、その理由はほとんどの場合、次のいずれかです。

- マルチパスデバイス選択の設定に整合性がない
- 存在しないデバイス参照を使用している

18.15.1 デバイス選択の問題の理解

ブロックデバイスは、マルチパスマップの一部にするか、または直接使用(ファイルシステムとしてマウント、スワップとして使用、LVM物理モジュール、その他)することができます。デバイスがすでにマウントされている場合、これをmultipathdでマルチパスマップの一部にしようとすると失敗し、「デバイスまたはリソースがビジーです」(Device or resource busy)というエラーが表示されます。逆に、すでにマルチパスマップの一部になっているデバイスをsystemdがマウントしようとすると同じエラーが発生します。

起動中のストレージデバイスのアクティブ化は、systemd、udev、multipathdおよび他のいくつかのツール間の複雑なやりとりで処理されます。udevルールが中心的な役割を果たします。これらは、デバイスの使用方法を他のサブシステムに指示するデバイスのプロパティを設定します。マルチパス関連のudevルールは、マルチパス処理用に選択されたデバイスに対して次のプロパティを設定します。

```
SYSTEMD_READY=0  
DM_MULTIPATH_DEVICE_PATH=1
```

パーティションデバイスは、これらのプロパティをその親から継承します。

これらのプロパティが正しく設定されていない場合、一部のツールがこれらのプロパティに従っていない場合、またはプロパティが設定されるのが遅すぎる場合、multipathdと他のサブシステムとの間で競合状態が発生する可能性があります。競合に勝ち残るのは1つだけです。それ以外は、「デバイスまたはリソースがビジーです」(Device or resource busy)というエラーが表示されます。

このコンテキストにおける1つの問題は、LVM2スイートのツールがデフォルトではudevプロパティを評価しないことです。これらは、デバイスがマルチパスコンポーネントかどうかを判定する独自のロジックを使用しますが、このロジックは、システムの残りの部分のロジックと一致しない場合があります。この回避方法は、[18.14.3項 「マルチパスデバイスでのLVM2の使用」](#)で説明されています。



注記: 起動デッドロックの例

ルートデバイスがマルチパス処理されておらず、マルチパスから除外されているデバイスがない、マルチパス処理を備えたシステムを考えてみます(18.3.2.2項「ローカルディスクのルートファイルシステム」のinitramfsでマルチパスを無効にするを参照)。このルートファイルシステムはinitramfsにマウントされています。systemdはルートファイルシステムに切り換わり、multipathdが起動します。デバイスがすでにマウントされているため、multipathdは、そのデバイスのマルチパスマップをセットアップできません。ルートデバイスはblacklistで設定されていないため、マルチパスデバイスとみなされ、このデバイス用にSYSTEMD_READY=0が設定されます。

起動プロセスの後半で、システムは、/varや/homeなどの追加のファイルシステムをマウントしようとします。通常、これらのファイルシステムはルートファイルシステムと同じデバイス上にあり、デフォルトではルートファイルシステム自体のBTRFSサブボリュームとして配置されます。ただし、systemdは、SYSTEMD_READY=0のためにこれらをマウントできません。「デッドロック状態」: dm-multipathデバイスを作成できず、systemdの基礎となるデバイスはブロックされます。追加のファイルシステムをマウントできず、起動に失敗します。

この問題の解決策はすでに存在します。multipathdはこの状況を検出し、デバイスをsystemdにリリースします。これにより、ファイルシステムのマウントを続行できます。ただし、一般的な問題を理解することが重要です。より分かりにくい形で発生する可能性もあるからです。

18.15.2 デバイス参照の問題の理解

デバイス参照の問題の例は18.7.4.2項「initramfsの永続的なデバイス名」で示しました。通常、1つのデバイスノードを指すシンボリックリンクは複数存在します(18.12.4項「マルチパスマップの参照」を参照)。ただし、これらのリンクは常に存在するわけではありません。udevは、現在のudevルールに従ってリンクを作成します。たとえば、マルチパス処理をオフにすると、/dev/mapper/の下にあるマルチパスデバイス用のシンボリックリンクはなくなります。したがって、/dev/mapper/デバイスへの参照は失敗します。

このような参照はさまざまな場所に(特に/etc/fstabおよび/etc/crypttab、initramfs、またはカーネルコマンドラインにも)表示されます。

この問題を回避する最も安全な方法は、起動と起動の間で持続しない種類のデバイス参照またはシステム設定に依存する種類のデバイス参照の使用を避けることです。一般的に、ファイルシステム(およびスワップ領域のような類似のエンティティ)は、含む側のデバイスではなく、ファイルシステム自体のプロパティ(UUIDやラベルなど)によって参照することをお勧め

します。このような参照を使用できず、たとえば、`/etc/crypttab`で、デバイス参照が必要な場合は、オプションを注意深く評価する必要があります。たとえば、[18.12.4項 「マルチパスマップの参照」](#)では、最良のオプションは`/dev/disk/by-id/wwn-`リンクかもしれません。これは`multipath=off`でも機能するからです。

18.15.3 緊急モードでのトラブルシューティング手順

微妙に異なるエラー状況が多数あるため、ステップバイステップのリカバリガイドを提供することは不可能です。ただし、これまでのサブセクションから背景知識を得ていれば、マルチパス処理の問題が原因でシステムが緊急モードになった場合、その問題を理解できるはずです。デバッグを始める前に、次の質問を確認してください。

- マルチパスサービスは有効になっていますか。
- initramfsにマルチパスdracutモジュールが含まれていますか。
- ルートデバイスがマルチパスデバイスとして設定されていますか。設定されていない場合、[18.11.1項 「multipath.confのblacklistセクション」](#)で説明されているように、ルートデバイスはマルチパスから正しく除外されていますか。または、initramfsにマルチパスモジュールがないことに依存していますか([18.3.2.2項 「ローカルディスクのルートファイルシステム」](#)を参照)。
- システムが緊急モードになったのは、実際のルートファイルシステムに切り替わる前ですか、または切り替わった後ですか。

最後の質問の答えがわからない場合、サンプルのdracut緊急プロンプトを次に示します。これは、ルートを切り替える前に出力されます。

```
Generating "/run/initramfs/rdsosreport.txt"
Entering emergency mode. Exit the shell to continue.
Type "journalctl" to view system logs.

You might want to save "/run/initramfs/rdsosreport.txt" to a USB stick or /boot
after mounting them and attach it to a bug report.

Give root password for maintenance
(or press Control-D to continue):
```

`rdsosreport.txt`が記載されているということは、システムがまだinitramfsから実行されていることを明確に示しています。それでも不明な場合は、ログインして`/etc/initrd-release`ファイルが存在するかチェックします。このファイルは、initramfs環境にのみ存在します。

ルートを切り替えた後に緊急モードになった場合、緊急プロンプトは同じように見えますが、rdsosreport.txtは記載されません。

```
Timed out waiting for device dev-disk-by\x2duuid-c4a...cfef77d.device.  
[DEPEND] Dependency failed for Local File Systems.  
[DEPEND] Dependency failed for Postfix Mail Transport Agent.  
Welcome to emergency shell  
Give root password for maintenance  
(or press Control-D to continue):
```

手順 18.2: 緊急モードの状況を分析する手順

- 障害の発生したsystemdユニットとジャーナルを調べて、障害が発生した原因を探ります。

```
# systemctl --failed  
# journalctl -b -o short-monotonic
```

ジャーナルを確認して、「最初に」障害が発生したユニットを特定します。最初の障害を見つけたら、その時点の前後のメッセージを注意深く調べます。警告またはその他の疑わしいメッセージはありますか。

ルートスイッチ(「Switching root.」)、およびSCSIデバイス、デバイスマッパー、マルチパスおよびLVM2に関するメッセージに注意します。デバイスおよびファイルシステムに関するsystemdメッセージ(「Found device...」、「Mounting...」、「Mounted...」)を探します。

- 既存のデバイス(低レベルのデバイスとデバイスマッパーテーバイスの両方)を調べます(以下のコマンドの一部はinitramfsでは使用できない場合があることに注意してください)。

```
# cat /proc/partitions  
# ls -l /sys/class/block  
# ls -l /dev/disk/by-id/* /dev/mapper/*  
# dmsetup ls --tree  
# lsblk  
# lsscsi
```

上記のコマンドの出力から、低レベルのデバイスが正常に検出されたかどうか、マルチパスマップおよびマルチパスパーティションがセットアップされたかどうかがわかるはずです。

- デバイスマッパーのマルチパスセットアップが予期どおりでない場合、udevプロパティ(特に、SYSTEMD_READY)を調べます(上記を参照)。

```
# udevadm info -e
```

4. 前の手順で予期しなかったudevプロパティが表示された場合、udevルールの処理中に不具合が発生した可能性があります。その他のプロパティ(特に、デバイスの識別に使用されたプロパティ)を確認します([18.12.1項 「WWIDとデバイスの識別」](#)を参照)。udevプロパティが正しい場合、ジャーナルでmultipathdのメッセージを再度確認します。「Device or resource busy」メッセージを探します。
5. システムがデバイスのマウントまたはアクティブ化に失敗した場合は、このデバイスを手動でアクティブ化してみると功を奏すことがあります。

```
# mount /var  
# swapon -a  
# vgchange -a y
```

ほとんどの場合、手動のアクティブ化は成功し、(通常は緊急シェルからログアウトするだけで)システム起動を続行し、起動したシステムでさらに状況を調べることができます。

手動のアクティブ化に失敗すると、多くの場合、不具合のヒントを示すエラーメッセージが表示されます。詳細度を上げてコマンドを再試行することもできます。

6. この時点で、問題は何であるかがある程度わかるはずです(そうでない場合は、SUSEサポートに連絡して、上記の質問のほとんどに回答できるように準備しておいてください)。いくつかのシェルコマンドで状況を修正し、緊急シェルを終了し、正常に起動できるはずです。同じ問題が今後発生しないよう、引き続き設定を調整する必要があります。または、レスキューシステムを起動し、デバイスを手動でセットアップして、chrootで実際のルートファイルシステムに変更し、前の手順で得た洞察に基づいて問題の修正を試みる必要があります。この状況では、ルートファイルシステムのストレージスタックが通常と異なっている場合があることに注意してください。セットアップに応じて、新しいinitramfsを構築するときにdracutモジュールの強制追加または省略が必要になる場合があります。[18.7.4.1項 「initramfsでのマルチパス処理の有効化または無効化」](#)も参照してください。

7. 問題の発生頻度が高いまたは起動のたびに問題が発生する場合、障害の詳細情報を得るために、詳細度を上げて起動を試します。次のカーネルパラメータ、またはそれらの組み合わせが役立つことがあります。

```
udev.log-priority=debug ①  
systemd.log_level=debug ②  
scsi_mod.scsi_logging_level=020400 ③  
rd.debug ④
```

- ① [systemd-udevd](#)とudevルール処理のログレベルを上げます。
- ② [systemd](#)のログレベルを上げます。
- ③ カーネルのSCSIサブシステムのロギングレベルを上げます。
- ④ initramfsのスクリプトを追跡します。

また、特定のドライバのロギングを有効にし、シリアルコンソールを設定して起動中の出力をキャプチャすることが役立つ場合があります。

18.15.4 技術情報ドキュメント

SUSE Linux Enterprise ServerのマルチパスI/Oの問題のトラブルシューティングの詳細については、SUSEナレッジベースにある、次のTID (技術情報ドキュメント)を参照してください。

- Using LVM on local and SAN attached devices (<https://www.suse.com/support/kb/doc/?id=000016331>) ↗
- Using LVM on Multipath (DM MPIO) Devices (<https://www.suse.com/support/kb/doc/?id=000017521>) ↗
- HOWTO: Add, Resize and Remove LUN without restarting SLES (<https://www.suse.com/support/kb/doc/?id=000017762>) ↗

19 NFS共有ファイルシステム

ネットワークファイルシステム(NFS)は、ローカルファイルへのアクセスと非常によく似た方法で、サーバ上のファイルにアクセスできるプロトコルです。

SUSE Linux Enterprise Serverは、NFS v4.2をインストールし、これにより、スパースファイル、ファイルの事前割り当て、サーバ側のクローンとコピー、アプリケーションデータブロック(ADB)、および必須アクセス制御(MAC)用のラベル付きNFS(クライアントとサーバの両方でMACが必要)のサポートが導入されます。

19.1 概要

「ネットワークファイルシステム」(NFS)は、標準化された、実証済みで幅広くサポートされているネットワークプロトコルであり、ファイルを別々のホスト間で共有することができます。

「ネットワーク情報サービス」(NIS)は、ネットワーク内で一元的なユーザ管理を行うために使用できます。NFSとNISを組み合わせることで、ネットワーク内のアクセス制御にファイルとディレクトリのパーミッションを使用できます。NFSをNISと連携して使用すると、ネットワークをユーザに対して透過的にすることができます。

デフォルト設定では、NFSはネットワークを完全に信頼しているので、信頼されたネットワークに接続されているマシンもすべて信頼します。NFSサーバが信頼するネットワークに物理的にアクセスできるコンピュータ上で管理者特権を持つユーザは、そのサーバが提供するファイルにアクセスできます。

多くの場合、このレベルのセキュリティは完全に満足のいくものであり(信頼されているネットワークが本当にプライベートである場合など)、しばしば単一のキャビネットや機械室に合わせてローカライズされており、不正なアクセスは不可能です。他のケースでは、1つのサブネット全体を1つの単位として信頼する必要性が制約となっており、よりきめの細かい信頼が求められます。これらのケースにおける必要性を満たすために、NFSは**Kerberos**インフラストラクチャを使用して、さまざまなセキュリティレベルをサポートしています。Kerberosには、デフォルトで使用されるNFSv4が必要です。詳細については、『Security and Hardening Guide』、第6章「Network authentication with Kerberos」を参照してください。

以下の用語は、YaSTモジュールで使用されています。

エクスポート

NFSサーバによって「エクスポートされ」、クライアントがシステムに統合できるディレクトリ。

NFSクライアント

NFSクライアントは、ネットワークファイルシステムプロトコルを介してNFSサーバからのNFSサービスを使用するシステムです。TCP/IPプロトコルはLinuxカーネルにすでに統合されており、追加ソフトウェアをインストールする必要はありません。

NFSサーバ

NFSサーバは、NFSサービスをクライアントに提供します。実行中のサーバは、次のデーモンに依存します。nfsd (ワーカ)、idmapd (NFSv4でのIDと名前のマッピング、特定のシナリオでのみ必要)、statd (ファイルのロック)、およびmountd (マウント要求)。

NFSv3

NFSv3はバージョン3の実装で、クライアント認証をサポートする「古い」ステートレスなNFSです。

NFSv4

NFSv4は、Kerberosによるセキュアなユーザ認証をサポートする新しいバージョン4の実装です。NFSv4で必要なポートは1つのみであるため、NFSv3よりもファイアウォール環境に適しています。

プロトコルは<https://datatracker.ietf.org/doc/rfc7531/>で指定されています。

pNFS

パラレル NFS。NFSv4のプロトコル拡張。任意のpNFSクライアントは、NFSサーバ上のデータに直接アクセスできます。

! 重要: DNSの必要性

原則として、すべてのエクスポートはIPアドレスのみを使用して実行できます。タイムアウトを回避するには、機能するDNSシステムが必要です。mountd デーモンは逆引きを行うので、少なくともログ目的でDNSは必要です。

19.2 NFSサーバのインストール

NFSサーバは、デフォルトインストールには含まれません。YaSTを使用してNFSサーバをインストールするには、ソフトウェア › ソフトウェア管理の順に選択し、パターンを選択して、Server Functions (サーバ機能)セクションでファイルサーバオプションを有効にします。了解をクリックして、必要なパッケージをインストールします。

このパターンには、NFSサーバのYaSTモジュールは含まれていません。パターンのインストールが完了した後、次のコマンドを実行してモジュールをインストールします。

```
> sudo zypper in yast2-nfs-server
```

NIS同様、NFSはクライアント/サーバシステムです。ただし、ファイルシステムをネットワーク経由で提供し(エクスポート)、同時に他のホストからファイルシステムをマウントすることができます(インポート)。



注記: NFSボリュームをエクスポート元サーバにローカルでマウントする

NFSボリュームのエクスポート元サーバへのローカルでのマウントは、SUSE Linux Enterprise Serverではサポートされていません。

19.3 NFSサーバの設定

NFSサーバの設定は、YaSTを使用するか、または手動で完了できます。認証のため、NFSをKerberosと組み合わせることもできます。

19.3.1 YaSTによるファイルシステムのエクスポート

YaSTを使用して、ネットワーク上のホストをNFSサーバにすることができます。NFSサーバとは、アクセスを許可されたすべてのホスト、またはグループのすべてのメンバーに、ディレクトリやファイルをエクスポートするサーバのことです。これにより、サーバは、ホストごとにアプリケーションをローカルインストールせずにアプリケーションを提供することもできます。

そのようなサーバをセットアップするには、次の手順に従います。

手順 19.1: NFSサーバをセットアップする

1. YaSTを起動し、ネットワークサービス › NFSサーバの順に選択します([図19.1 「NFSサーバ設定ツール」](#)を参照してください)。追加のソフトウェアをインストールするよう求められることがあります。



図 19.1: NFS サーバ設定ツール

2. 開始ラジオボタンをクリックします。
3. firewalldがシステムでアクティブな場合は、NFS用に個別に設定します(19.5項「ファイアウォールの背後でNFSサーバとクライアントを操作する」を参照)。YaSTはまだ、firewalldを完全にはサポートしていないため、「ファイアウォールを設定できません」というメッセージを無視して続行します。
4. NFSv4を有効にするを選択するかどうかを決定します。NFSv4を無効にした場合、YaSTでサポートされるのはNFSv3のみになります。NFSv2の有効化の詳細については、注記:NFSv2を参照してください。
 - NFSv4を選択した場合は、追加で適切なNFSv4ドメイン名を入力します。このパラメータは、Kerberosの設定に必要なidmapdデーモンによって使用されるか、クライアントが数字のユーザ名を処理できない場合に使用されます。idmapdを実行しない場合、または特に必要のない場合は、そのままlocaldomain(デフォルト)を使用してください。idmapdデーモンの詳細については、/etc/idmapd.confを参照してください。



重要: NFSv4 ドメイン名

ドメイン名はすべてのNFSv4クライアントで同様に設定する必要があることに注意してください。サーバと同じドメイン名を共有するクライアントのみがサーバにアクセスできます。サーバとクライアントのデフォルトのドメイン名は`localdomain`です。

5. サーバに安全にアクセスするには、GSSセキュリティを有効にするをクリックします。この手順の前提条件として、ドメインにKerberosをインストールし、サーバとクライアントの両方でKerberosを有効にしておく必要があります。次へをクリックして、次の設定ダイアログに進みます。
6. ディレクトリをエクスポートするには、ダイアログの上半分にあるディレクトリの追加をクリックします。
7. 許可されるホストをまだ設定していない場合は、自動的に別のダイアログが表示されるので、クライアント情報およびオプションを入力します。ホストを示すワイルドカードを入力します(通常はデフォルト設定のまま使用できます)。
4種類の方法でホストを指定することができます。1台のホスト(名前またはIPアドレス)(single host)、ネットグループ(netgroups)、ワイルドカード(すべてのマシンがサーバにアクセスできることを示す*など)(wild cards)、およびIPネットワーク(IP networks)です。
これらのオプションの詳細については、[exports](#)のマニュアルページを参照してください。
8. 完了をクリックして設定を完了します。

19.3.2 ファイルシステムの手動エクスポート

NFSエクスポートサービスの環境設定ファイルは、`/etc(exports`と`/etc/sysconfig/nfs`です。Kerberized NFSを使用したNFSv4サーバ設定に必要な場合、またはクライアントが数字のユーザ名を処理できない場合は、これらのファイル以外に`/etc/idmapd.conf`も必要です。サービスを起動または再起動するには、`sudo systemctl restart nfs-server`を実行します。これにより、NFSサーバで必要なRPCポートマッパーも再起動されます。NFSサーバがブート時に常に起動するようにするには、`sudo systemctl enable nfs-server`を実行します。



注記: NFSv4

NFSv4は、SUSE Linux Enterprise Serverで利用できる最新版のNFSプロトコルです。NFSv3と同じ方法で、NFSv4でのエクスポート用にディレクトリを設定できるようになりました。

SUSE Linux Enterprise Server 11では、`/etc(exports`のバインドマウントが必須でした。これは引き続きサポートされていますが、非推奨になりました。

/etc(exports

/etc(exportsファイルには、エントリのリストが含まれています。各エントリはそれぞれ共有するディレクトリと共有方法を示します。/etc(exports中の一般的なエントリは、次の項目から成り立っています。

```
/SHARED/DIRECTORY    HOST(OPTION_LIST)
```

例:

```
/nfs_exports/public *(rw,sync,root_squash,wdelay)
/nfs_exports/department1 *.department1.example.com(rw,sync,root_squash,wdelay)
/nfs_exports/team1 192.168.1.0/24(rw,sync,root_squash,wdelay)
/nfs_exports/tux 192.168.1.2(rw,sync,root_squash)
```

この例では、HOSTの次の値が使用されています。

- *: ネットワーク上のすべてのクライアントにエクスポートします
- *.department1.example.com: *.department1.example.comドメインのクライアントにのみエクスポートします
- 192.168.1.0/24: 192.168.1.0/24の範囲のIPアドレスを持つクライアントにのみエクスポートします
- 192.168.1.2: IPアドレスが192.168.1.2のマシンにのみエクスポートします

上記の例に加えて、/etc/netgroupで定義されたネットグループ(@my-hosts)へのエクスポートも制限することができます。各オプションの詳細とその意味については、/etc(exportsのmanページ(man exports)を参照してください。

NFSサーバの実行中に/etc(exportsを変更した場合、変更を有効にするには、sudo systemctl restart nfs-serverを実行してサーバを再起動する必要があります。

/etc/sysconfig/nfs

/etc/sysconfig/nfs ファイルには、 NFSv4 サーバデーモンの動作を決定する小数のパラメータが含まれています。 NFS4_SUPPORT パラメータを yes に設定することが重要です (デフォルトの設定)。 NFS4_SUPPORT は、 NFS サーバが NFSv4 エクスポートとクライアントをサポートするかどうかを決定します。

NFS サーバの実行中に /etc/sysconfig/nfs を変更した場合、 変更を有効にするには、 sudo systemctl restart nfs-server を実行してサーバを再起動する必要があります。



ヒント: マウントオプション

SUSE Linux Enterprise Server 11 では、 /etc(exports) の --bind マウントが必須でした。これは引き続きサポートされていますが、 非推奨になりました。 NFSv3 と同じ方法で、 NFSv4 でのエクスポート用にディレクトリを設定できるようになりました。



注記: NFSv2

NFS クライアントがまだ NFSv2 に依存している場合は、 サーバの /etc/sysconfig/nfs に次のように設定して NFSv2 を有効にします。

```
NFSD_OPTIONS="-V2"  
MOUNTD_OPTIONS="-V2"
```

サービスを再起動した後で、 次のコマンドを実行して、 バージョン2が使用可能かどうかを確認します。

```
> cat /proc/fs/nfsd/versions  
+2 +3 +4 +4.1 +4.2
```

/etc/idmapd.conf

idmapd デーモンは、 Kerberos 認証を使用する場合、 またはクライアントが数字のユーザ名を処理できない場合にのみ必要です。 Linux クライアントは、 Linux カーネル 2.6.39 から数字のユーザ名を処理できるようになりました。idmapd デーモンは、 NFSv4 からサーバへの要求に対して名前と ID のマッピングを行い、 クライアントに応答します。

必要に応じて、 idmapd を NFSv4 サーバ上で実行する必要があります。 クライアントの名前と ID のマッピングは、 パッケージ nfs-client によって提供される nfsidmap によって行われます。

NFSを使ってファイルシステムを共有するマシン間では、ユーザへのユーザ名とID (UID) の割り当てには同じ方法を使用してください。そのためには、NIS、LDAP、または他の同一ドメイン認証機構を利用することができます。

/etc/idmapd.confのDomainパラメータを設定する必要があります。このパラメータはサーバとこのサーバにアクセスするすべてのNFSv4クライアントに対して同じである必要があります。異なるNFSv4ドメインのクライアントはサーバにアクセスできません。デフォルトのドメインlocaldomainをそのまま使用することをお勧めします。別の名前を選択する必要がある場合は、ホストのFQDNからホスト名を除いたものにするとよいでしょう。環境設定ファイルの例を次に示します。

```
[General]
Verbosity = 0
Pipes-Directory = /var/lib/nfs/rpc_pipefs
Domain = localdomain

[Mapping]
Nobody-User = nobody
Nobody-Group = nobody
```

idmapdデーモンを起動するため、systemctl start nfs-idmapdを実行します。デーモンの実行中に/etc/idmapd.confを変更した場合、変更を有効にするには、systemctl restart nfs-idmapdを実行してデーモンを再起動する必要があります。

詳細については、idmapdおよびidmapd.confのマニュアルページ(man idmapdおよびman idmapd.conf)を参照してください。

19.3.3 NFSでのKerberosの使用

NFSでKerberos認証を使用するには、Generic Security Services (GSS)を有効にする必要があります。最初のYaST NFSサーバのダイアログで、GSSセキュリティを有効にするを選択します。ただし、この機能を使用するには、機能するKerberosサーバが必要です。YaSTはKerberosサーバの設定は行いません。その提供機能を使用するだけです。YaST環境設定に加えて、Kerberos認証も使用するには、NFS設定を実行する前に、少なくとも次の手順を完了してください。

1. サーバとクライアントの両方が、同じKerberosドメインにあることを確認します。つまり、クライアントとサーバが同じKDC (Key Distribution Center)サーバにアクセスし、krb5.keytabファイル(マシン上のデフォルトの場所は/etc/krb5.keytab)を共有していかなければなりません。Kerberosの詳細については、『Security and Hardening Guide』、第6章「Network authentication with Kerberos」を参照してください。

2. クライアントで`systemctl start rpc-gssd.service`コマンドを実行して、gssdサービスを起動します。
3. サーバで`systemctl start rpc-svcgssd.service`コマンドを実行して、svcgssdサービスを起動します。

Kerberos認証でも、サーバで`idmapd`デーモンが実行されている必要があります。詳細については、[/etc/idmapd.conf](#)を参照してください。

Kerberos化されたNFSの設定の詳細については、[19.7項 「詳細情報」](#)のリンクを参照してください。

19.4 クライアントの設定

ホストをNFSクライアントとして設定する場合、他のソフトウェアをインストールする必要はありません。必要なすべてのパッケージは、デフォルトでインストールされます。

19.4.1 YaSTによるファイルシステムのインポート

認証されたユーザは、YaST NFSクライアントモジュールを使用して、NFSディレクトリをNFSサーバからローカルファイルツリーにマウントできます。以下に手順を示します。

手順 19.2: NFSディレクトリのインポート

1. YaST NFSクライアントモジュールを起動します。
2. NFS共有タブで追加をクリックします。NFSサーバのホスト名、インポートするディレクトリ、およびこのディレクトリをローカルでマウントするマウントポイントを入力します。
3. NFSv4を使用する場合は、NFS設定タブでNFSv4を有効にするを選択します。また、NFSv4ドメイン名に、NFSv4サーバが使用する値と同じ値が入力されている必要があります。デフォルトドメインは`localdomain`です。
4. NFSでKerberos認証を使用するには、GSSセキュリティを有効にする必要があります。GSSセキュリティを有効にするを選択します。
5. `firewalld`がシステムでアクティブな場合は、NFS用に個別に設定します([19.5項 「ファイアウォールの背後でNFSサーバとクライアントを操作する」](#)を参照)。YaSTはまだ、`firewalld`を完全にはサポートしていないため、「ファイアウォールを設定できません」というメッセージを無視して続行します。

6. OKをクリックして変更内容を保存します。

設定は`/etc/fstab`に書かれ、指定されたファイルシステムがマウントされます。後でYaST設定クライアントを起動した時に、このファイルから既存の設定が取得されます。



ヒント: ルートファイルシステムとしてのNFS

ルートパーティションがネットワーク経由でNFS共有としてマウントされている(ディスクレス)システムでは、NFS共有にアクセス可能なネットワークデバイスの設定を慎重に行う必要があります。

システムの停止、システムの再起動時のデフォルトの処理順序は、ネットワーク接続を切断してから、ルートパーティションをアンマウントするという順序になります。NFSルートの場合、この順序では問題が発生します。NFS共有とのネットワーク接続が先に無効にされているため、ルートパーティションを正常にアンマウントできないためです。システムが該当するネットワークデバイスを無効にしないようにするには、[network device configuration(ネットワークデバイスの設定)]タブ(『管理ガイド』、第23章「ネットワークの基礎」、23.4.1.2.5項「ネットワークデバイスの有効化」を参照)を開いて、デバイスの起動ペインのNFSrootオンを選択します。

19.4.2 ファイルシステムの手動インポート

NFSサーバからファイルシステムを手動でインポートするには、RPCポートマッパーが実行していることが前提条件です。RPCポートマッパーを適切に起動するのは`nfs`サービスです。そのため、`root`ユーザとして「`sudo systemctl start nfs`」を入力し、RPCポートマッパーを起動します。次に、`mount`を使用して、ローカルパーティションと同様に、リモートファイルシステムをファイルシステムにマウントできます。

```
> sudo mount HOST:REMOTE-PATH LOCAL-PATH
```

たとえば、`nfs.example.com`マシンからユーザディレクトリをインポートするには、次の構文を使用します。

```
> sudo mount nfs.example.com:/home /home
```

クライアントがNFSサーバに対して行うTCP接続の数を定義するには、`nconnect`コマンドの`mount`オプションを使用できます。1~16の間の任意の数を指定できます。ここで、1はマウントオプションが指定されていない場合のデフォルト値です。

nconnect設定は、特定のNFSサーバへの最初のマウントプロセス中にのみ適用されます。同じクライアントが同じNFSサーバにmountコマンドを実行する場合、すべてのすでに確立されている接続が共有されます。新しい接続は確立されません。nconnect設定を変更するには、特定のNFSサーバへの「すべての」クライアント接続をアンマウントする必要があります。次にnconnectオプションの新しい値を定義できます。

現在有効なnconnectの値は、mountの出力または/proc/mountsファイルで確認できます。マウントオプションに値がない場合、そのオプションはマウント中に使用されず、デフォルト値の1が使用されます。



注記: nconnectによって定義されているものとは異なる接続数

最初のマウント後に接続を閉じたり、開いたりできるため、実際の接続数は、必ずしもnconnectの値と同じである必要はありません。

19.4.2.1 自動マウントサービスの使用

autofsデーモンを使用して、リモートファイルシステムを自動的にマウントすることができます。/etc/auto.masterファイルに次のエントリを追加します。

```
/nfsmounts /etc/auto.nfs
```

これで、/nfsmountsディレクトリがクライアント上のすべてのNFSマウントのルートディレクトリの役割を果たすようになります(auto.nfsファイルが正しく設定されている場合)。ここでは、auto.nfsという名前を使用しましたが、任意の名前を選択することができます。auto.nfsで、次のようにしてすべてのNFSマウントのエントリを追加します。

```
localdata -fstype=nfs server1:/data
nfs4mount -fstype=nfs4 server2:/
```

rootユーザとしてsystemctl start autofsを実行して設定を有効にします。この例では、server1の/dataディレクトリの/nfsmounts/localdataはNFSでマウントされ、server2の/nfsmounts/nfs4mountはNFSv4でマウントされます。

autofsサービスの実行中に/etc/auto.masterファイルを編集した場合、変更を反映するには、systemctl restart autofsで自動マウント機能を再起動する必要があります。

19.4.2.2 /etc/fstabの手動編集

/etc/fstab内の典型的なNFSv3マウントエントリは、次のようになります。

```
nfs.example.com:/data /local/path nfs rw,noauto 0 0
```

NFSv4マウントの場合は、3番目の列でnfsの代わりにnfs4を使用します。

```
nfs.example.com:/data /local/pathv4 nfs4 rw,noauto 0 0
```

noautoオプションを使用すると、起動時にファイルシステムが自動マウントされません。対応するファイルシステムを手動でマウントする場合は、マウントポイントのみを指定してmountコマンドを短くできます。

```
> sudo mount /local/path
```



注記: 起動時にマウント

ただし、noautoオプションを入力しないと、起動時に、システムのinitスクリプトによって、それらのファイルシステムがマウントされます。その場合は、ネットワークが利用可能になる前にスクリプトが共有をマウントしようとするのを防ぐ、_netdevオプションを追加することを検討してください。

19.4.3 パラレルNFS(pNFS)

NFSは、1980年代に開発された、もっとも古いプロトコルの1つです。そのため、小さなファイルを共有したい場合は、通常、NFSで十分です。しかし、大きなファイルを送信したい場合や多数のクライアントがデータにアクセスしたい場合は、NFSサーバがボトルネックとなり、システムのパフォーマンスに重大な影響を及ぼします。これはファイルのサイズが急速に大きくなっているのに対し、Ethernetの相対速度が追い付いていないためです。

通常のNFSサーバにファイルを要求すると、サーバはファイルのメタデータを検索し、すべてのデータを収集して、ネットワークを介してクライアントに送信します。しかし、ファイルが小さくても大きくてもパフォーマンスのボトルネックが問題になります。

- 小さいファイルでは、メタデータの収集に時間がかかる。
- 大きいファイルでは、サーバからクライアントへのデータ送信に時間がかかる。

pNFS(パラレルNFS)は、ファイルシステムメタデータをデータの場所から分離することによって、この制限を克服します。このため、pNFSには2種類のサーバが必要です。

- データ以外のすべてのトランザクションを扱う「メタデータ」または「制御サーバ」
- データを保持する1つ以上の「ストレージサーバ」

メタデータサーバとストレージサーバによって、単一の論理NFSサーバが構成されます。クライアントが読み込みまたは書き出しを行う場合、メタデータサーバがNFSv4クライアントに対して、ファイルのチャンクにアクセスするにはどのストレージサーバを使用すればよいかを指示します。クライアントはサーバのデータに直接アクセスできます。

SUSE Linux Enterprise Serverはクライアント側でのみpNFSをサポートします。

19.4.3.1 YaSTを使用したpNFSクライアントの設定

手順19.2 「NFSディレクトリのインポート」に従って進めます。ただし、pNFS (v4.2) チェックボックスをクリックし、オプションでNFSv4共有をクリックします。YaSTが必要な手順をすべて実行し、必要なすべてのオプションを /etc(exports) ファイルに書き込みます。

19.4.3.2 pNFSクライアントの手動設定

19.4.2項 「ファイルシステムの手動インポート」を参照して開始します。ほとんどの設定は NFSv4サーバによって行われます。pNFSを使用する場合に異なるのは、nfsvers オプションおよびメタデータサーバMDS_SERVERを mount コマンドに追加することだけです。

```
> sudo mount -t nfs4 -o nfsvers=4.2 MDS_SERVER MOUNTPOINT
```

デバッグを支援するために、/proc ファイルシステムの値を変更します。

```
> sudo echo 32767 > /proc/sys/sunrpc/nfsd_debug  
> sudo echo 32767 > /proc/sys/sunrpc/nfs_debug
```

19.5 ファイアウォールの背後でNFSサーバとクライアントを操作する

NFSサーバとそのクライアントとの間の通信は、リモートプロシージャコール(RPC)を介して行われます。マウントデーモンやファイルロックサービスなど、いくつかのRPCサービスは、Linux NFS実装の一部です。サーバとクライアントがファイアウォールの背後で実行される場合、これらのサービスとファイアウォールは、クライアントとサーバ間の通信をブロックしないように設定する必要があります。

NFS 4サーバはNFSバージョン3と後方互換性がありますが、ファイアウォール設定は両方のバージョンで異なります。クライアントのいずれかがNFS 3を使用して共有をマウントする場合は、NFS 4とNFS 3の両方を許可するようにファイアウォールを設定します。

19.5.1 NFS 4.x

NFS 4では、サーバ側でのみTCPポート2049を開く必要があります。ファイアウォールでこのポートを開くには、「NFSサーバの」firewalldでnfsサービスを有効にします。

```
> sudo firewall-cmd --permanent --add-service=nfs --zone=ACTIVE_ZONE  
firewall-cmd --reload
```

ACTIVE_ZONEを、NFSサーバで使用されるファイアウォールゾーンに置き換えます。

NFSv4の使用時には、クライアント側での追加のファイアウォール設定は必要ありません。デフォルトでは、マウントのデフォルトはサポートされているNFSの最高バージョンとなるため、クライアントがNFSv4をサポートしている場合、共有は自動的にバージョン4.2としてマウントされます。

19.5.2 NFS 3

NFS 3では次のサービスが必要です。

- portmapper
- nfsd
- mountd
- lockd
- statd

これらのサービスは、rpcbindによって操作され、デフォルトでは動的にポートが割り当てられます。ファイアウォールの背後でこれらのサービスにアクセスできるようにするには、まず、静的ポートで実行するように設定する必要があります。これらのポートは後でファイアウォールで開く必要があります。

portmapper

SUSE Linux Enterprise Serverでは、portmapperはすでに静的ポートで実行するように設定されています。

ポート	111
プロトコル	TCP、 UDP
実行環境	クライアント、 サーバ

```
> sudo firewall-cmd --add-service=rpc-bind --permanent --zone=ACTIVE_ZONE
```

nfsd

SUSE Linux Enterprise Serverでは、nfsdはすでに静的ポートで実行するように設定されています。

ポート	2049
プロトコル	TCP、 UDP
実行環境	サーバ

```
> sudo firewall-cmd --add-service=nfs3 --permanent --zone=ACTIVE_ZONE
```

mountd

SUSE Linux Enterprise Serverでは、mountdはすでに静的ポートで実行するように設定されています。

ポート	20048
プロトコル	TCP、 UDP
実行環境	サーバ

```
> sudo firewall-cmd --add-service=mountd --permanent --zone=ACTIVE_ZONE
```

lockd

lockdの静的ポートを設定するには、次の操作を実行します。

1. サーバ上の/etc/sysconfig/nfsを編集し、以下を見つけて設定します

```
LOCKD_TCPPORT=NNNN  
LOCKD_UDPPORT=NNNN
```

NNNNを、任意の未使用ポートに置き換えます。両方のプロトコルに同じポートを使用します。

2. NFSサーバを再起動します。

```
> sudo systemctl restart nfs-server
```

ポート	NNNN
-----	------

プロトコル	TCP、 UDP
実行環境	クライアント、 サーバ
> sudo firewall-cmd --add-port=NNNNN/{tcp,udp} --permanent --zone=ACTIVE_ZONE	

statd

statdの静的ポートを設定するには、次の操作を実行します。

1. サーバ上の/etc/sysconfig/nfsを編集し、以下を検出して設定します

```
STATD_PORT=NNNNN
```

NNNNNを、任意の未使用ポートに置き換えます。

2. NFSサーバを再起動します。

```
> sudo systemctl restart nfs-server
```

ポート	NNNNN
プロトコル	TCP、 UDP
実行環境	クライアント、 サーバ
> sudo firewall-cmd --add-port=NNNNN/{tcp,udp} --permanent --zone=ACTIVE_ZONE	



重要: 変更されたfirewalld設定のロード

firewalld設定を変更するたびに、デーモンを再ロードして、変更を有効にする必要があります。

```
> sudo firewall-cmd --reload
```



注記: ファイアウォールゾーン

ACTIVE_ZONEを、各マシンで使用されているファイアウォールゾーンに置き換えてください。ファイアウォール設定によっては、アクティブゾーンがマシンごとに異なる場合があることに注意してください。

19.6 NFSv4上でのアクセス制御リストの管理

Linuxには、ユーザ、グループ、およびその他(rwx)に対する簡単な読み込み、書き込み、および実行(ugo)の各フラグ以上の、ACL (アクセス制御リスト)の単一標準はありません。よりきめ細かな制御のオプションの1つにDraft POSIX ACLsがあります。ただし、これらのACLは、POSIXによって正式に標準化されたことはありません。もう1つは、NFSv4ネットワークファイルシステムの一部として設計されたNFSv4 ACLです。NFSv4 ACLは、Linux上のPOSIXシステムとMicrosoft Windows上のWIN32システム間に適切な互換性を提供することを目的としています。

NFSv4 ACLは、Draft POSIX ACLを正しく実装できるほど十分ではないので、NFSv4クライアントへのACLアクセスのマッピングは試みられていません(setfaclの使用など)。

NFSv4の使用時は、Draft POSIX ACLはエミュレーションでさえ使用できず、NFSv4 ACLを直接使用する必要があります。つまり、setfaclをNFSv3で動作させながら、NFSv4で動作させることはできません。NFSv4 ACLをNFSv4ファイルシステムで使用できるようにするために、SUSE Linux Enterprise Serverでは、次のファイルを含むnfs4-acl-toolsパッケージを提供しています。

- nfs4-getfacl
- nfs4-setfacl
- nfs4-editacl

これらの動作は、NFSv4 ACLを検証および変更するgetfaclおよびsetfaclとほぼ同様です。これらのコマンドは、NFSサーバ上のファイルシステムがNFSv4 ACLを完全にサポートしている場合にのみ有効です。サーバによって課される制限は、クライアントで実行されているこれらのプログラムに影響を与え、ACE (Access Control Entries)の一部の特定の組み合わせが不可能なことがあります。

エクスポート元のNFSサーバにNFSボリュームをローカルにマウントすることはサポートされていません。

その他の情報

詳細については、Introduction to NFSv4 ACLs (https://wiki.linux-nfs.org/wiki/index.php/ACLs#Introduction_to_NFSv4_ACLs)を参照してください。

19.7 詳細情報

exports、nfs、およびmountのマニュアルページに加えて、NFSサーバとクライアントの設定に関する情報は、/usr/share/doc/packages/nfsidmap/READMEで入手できます。オンラインドキュメンテーションについては、次のWebサイトを参照してください。

- ネットワークセキュリティの一般的な情報については、『Security and Hardening Guide』、第23章「Masquerading and firewalls」を参照してください。
- NFSエクスポートを自動的にマウントする必要がある場合は、21.4項「NFS共有の自動マウント」を参照してください。
- AutoYaSTを使用してNFSを設定する方法の詳細については、『AutoYaST Guide』、第4章「Configuration and installation options」、4.21項「NFS client and server」を参照してください。
- Kerberosを使用したNFSエクスポートのセキュリティ保護に関する手順については、『Security and Hardening Guide』、第6章「Network authentication with Kerberos」、6.6項「Kerberos and NFS」を参照してください。
- 詳細な技術ヘルプについては、SourceForge (<https://nfs.sourceforge.net/>) を参照してください。

19.8 NFSトラブルシューティングのための情報の収集

19.8.1 一般的なトラブルシューティング

生成されたエラーメッセージを読み、/var/log/messagesファイルを調べることで、NFSでの問題を理解できる場合があります。ただし、多くの場合、エラーメッセージや/var/log/messagesで提供される情報は、詳細が十分ではありません。このような場合、NFSのほとんどの問題は、問題の再現中にネットワークパケットをキャプチャすることでよく理解することができます。

問題を明確に定義します。さまざまな方法でシステムをテストし、問題がいつ発生するのかを特定して問題を調べます。問題につながる最も簡単なステップを特定します。その後、次の手順で示すように、問題を再現してみます。

手順 19.3: 問題の再現

- ネットワークパケットをキャプチャします。Linuxでは、tcpdumpパッケージで提供される、tcpdumpコマンドを使用できます。

tcpdumpの構文の例は次のとおりです。

```
tcpdump -s0 -i eth0 -w /tmp/nfs-demo.cap host x.x.x.x
```

ここで、

s0

パケットの切り捨てを防止します

eth0

パケットが通過するローカルインターフェースの名前に置き換える必要があります。any値を使用して、同時にすべてのインターフェースをキャプチャできますが、この属性の使用により、データが劣化したり、分析で混乱が生じる場合がよくあります。

w

書き込むキャプチャファイルの名前を指定します。

x.x.x.x

NFS接続のもう一方の端のIPアドレスに置き換える必要があります。たとえば、NFSクライアント側でtcpdumpを取得する場合は、NFSサーバのIPアドレスを指定します。その逆も同様です。



注記

場合によっては、NFSクライアントまたはNFSサーバのいずれかのデータをキャプチャするだけで十分です。ただし、エンドツウエンドのネットワーク整合性が疑わしい場合は、両方の端でデータをキャプチャする必要がよくあります。

tcpdumpプロセスをシャットダウンせずに次の手順に進みます。

- (オプション) nfs mountコマンド自体の実行中に問題が発生する場合は、nfs mountコマンドの高詳細度オプション(-vvv)を使用して、より詳細な出力を得ることができます。

3. (オプション)再現方法のstraceを取得します。straceの再現ステップでは、どのシステムコールがどの時点で行われたかを正確に記録します。この情報を使用して、tcpdump内のどのイベントに焦点を合わせるべきかを詳細に判断することができます。

たとえば、NFSマウントでコマンド **mycommand --param** の実行が失敗したことが分かった場合は、次のコマンドを使用してコマンドをstraceすることができます。

```
strace -ttf -s128 -o/tmp/nfs-strace.out mycommand --param
```

再現ステップでstraceを取得できない場合は、問題が再現された時刻を記録します。/var/log/messagesログファイルを確認して、問題を特定します。

4. 問題が再現されたら、**CTRL + c** を押して、端末で実行しているtcpdumpを停止します。straceコマンドによりハングした場合は、straceコマンドも終了します。

5. これで、パケットトレースとstraceデータの分析経験のある管理者は、/tmp/nfs-demo.capと/tmp/nfs-strace.outのデータを検査できるようになりました。

19.8.2 高度なNFSデバッグ

! 重要: 高度なデバッグは専門家向けです

次のセクションは、NFSコードを理解している熟練したNFS管理者のみを対象としていることを念頭に置いてください。したがって、19.8.1項「一般的なトラブルシューティング」で説明されている最初の手順を実行して、問題を絞り込み、詳細を得るために必要なデバッグコード(ある場合)の領域を専門家に知らせてください。

追加のNFS関連の情報を収集するために有効にすることが可能なデバッグコードのさまざまな領域があります。ただし、デバッグメッセージは非常に難解であり、その量も非常に多いため、デバッグコードの使用はシステムのパフォーマンスに影響を及ぼす可能性があります。問題が発生しないようにするためにシステムに大きな影響を及ぼす場合もあります。ほとんどの場合、デバッグコードの出力は必要ありません。また、NFSコードに精通していないユーザにとっては通常は役に立ちません。

19.8.2.1 **rpcdebug**を使用したデバッグの有効化

rpcdebugツールを使用すると、NFSクライアントとサーバのデバッグフラグを設定したり、クリアしたりできます。SUSE Linux Enterprise Serverのインストールで**rpcdebug**ツールが利用できない場合は、NFSサーバのnfs-clientあるいはnfs-kernel-serverパッケージからインストールできます。

デバッグフラグを設定するには、次のコマンドを実行します。

```
rpcdebug -m module -s flags
```

デバッグフラグをクリアするには、次のコマンドを実行します。

```
rpcdebug -m module -c flags
```

ここで、moduleは次のとおりです。

nfsd

NFSサーバコードのデバッグ

nfs

NFSクライアントコードのデバッグ

nlm

NFSクライアントまたはNFSサーバのいずれかでNFS Lock Managerのデバッグを行います。これはNFS v2/v3に該当します。

rpc

NFSクライアントまたはNFSサーバのいずれかでリモートプロシージャコールモジュールのデバッグを行います。

rpcdebugコマンドの詳細な使用法については、マニュアルページを参照してください。

```
man 8 rpcdebug
```

19.8.2.2 NFSが依存する他のコードのデバッグを有効化する

NFSアクティビティは、NFSマウントデーモン(**rpc.mountd**)などの他の関連サービスに依存する場合があります。/etc/sysconfig/nfs内で関連サービスのオプションを設定できます。

たとえば、/etc/sysconfig/nfsには次のパラメータが含まれています。

```
MOUNTD_OPTIONS=""
```

デバッグモードを有効にするには、`-d`オプションに続いて、`all`、`auth`、`call`、`general`、または`parse`のいずれかの値を使用する必要があります。

たとえば、次のコードはすべての形式のrpc.mountdロギングを有効にします。

```
MOUNTD_OPTIONS="-d all"
```

すべての使用可能なオプションについては、マニュアルページを参照してください。

```
man 8 rpc.mountd
```

/etc/sysconfig/nfsを変更した後で、サービスを再起動する必要があります。

```
systemctl restart nfs-server # for nfs server related changes  
systemctl restart nfs # for nfs client related changes
```

20 Samba

Sambaを使用すると、macOS、Windows、OS/2マシンに対するファイルサーバおよびプリントサーバをUnixマシン上に構築できます。Sambaは、今や成熟の域に達したかなり複雑な製品です。YaSTで、または環境設定ファイルを手動で編集することで、Sambaを設定します。

! 重要: SMB1はサポートされていない

Sambaバージョン4.17以降、SUSE Linux Enterprise ServerではSMB1プロトコルが無効になっており、サポートされなくなりました。

20.1 用語集

ここでは、SambaのマニュアルやYaSTモジュールで使用される用語について説明します。

SMBプロトコル

SambaはSMB(サーバメッセージブロック)プロトコルを使用します。SMBはNetBIOSサービスを基にしています。Microsoftは、他のメーカーのソフトウェアがMicrosoftオペレーティングシステムを実行しているサーバへの接続を確立できるよう、このプロトコルをリリースしました。SambaはTCP/IPプロトコルの上にSMBプロトコルを実装します。つまり、すべてのクライアントにTCP/IPをインストールし、有効にする必要があります。



ヒント: IBM Z: NetBIOSのサポート

IBM ZではSMB over TCP/IPのみがサポートされています。これら2つのシステムではNetBIOSをサポートしていません。

CIFSプロトコル

CIFS (Common Internet File System) プロトコルは、SMB1とも呼ばれるSMBプロトコルの初期バージョンです。CIFSは、TCP/IP上で使用する標準のリモートファイルシステムで、ユーザグループによる共同作業およびインターネットを介したドキュメントの共有ができるようにします。

SMB1はSMB2に置き換えられ、Microsoft Windows Vista™の一部として最初にリリースされました。これは、Microsoft Windows 8™およびMicrosoft Windows Server 2012ではSMB3に置き換えられました。最新バージョンのSambaでは、セキュリティ上の理由によりデフォルトでSMB1は無効になっています。

NetBIOS

NetBIOSは、ネットワーク上のコンピュータ間の名前解決と通信用に設計されたソフトウェアインターフェース(API)です。これにより、ネットワークに接続されたマシンが、それ自身の名前を維持できます。予約を行えば、これらのマシンを名前によって指定できます。名前を確認する一元的なプロセスはありません。ネットワーク上のマシンでは、すでに使用済みの名前でない限り、名前をいくつでも予約できます。NetBIOSはさまざまなネットワークプロトコルの上に実装できます。比較的単純でルーティング不可能な実装の1つは、NetBEUIと呼ばれます(これはNetBIOS APIと混同されることが多くあります)。NetBIOSは、Novell IPX/SPXプロトコルの上でもサポートされています。バージョン3.2以降、SambaはIPv4とIPv6の両方でNetBIOSをサポートしています。

TCP/IP経由で送信されたNetBIOS名は、/etc/hostsで使用されている名前、またはDNSで定義された名前とまったく共通点がありません。NetBIOSは独自の、完全に独立した名前付け規則を使用しています。しかし、管理を容易にするために、またはDNSをネイティブで使用するために、DNSホスト名に対応する名前を使用することをお勧めします。これはSambaが使用するデフォルトでもあります。

Sambaサーバ

Sambaサーバは、SMB/CIFSサービスおよびNetBIOS over IPネーミングサービスをクライアントに提供します。Linuxの場合、3種類のSambaサーバデーモン(SMB/CIFSサービス用smbd、ネーミングサービス用nmbd、認証用winbind)が用意されています。

Sambaクライアント

Sambaクライアントは、SMBプロトコルを介してSambaサーバからSambaサービスを使用するシステムです。WindowsやmacOSなどの一般的なオペレーティングシステムは、SMBプロトコルをサポートしています。TCP/IPプロトコルは、すべてのコンピュータにインストールする必要があります。Sambaは、異なるUNIXフレーバーに対してクライアントを提供します。Linuxでは、SMB用のカーネルモジュールがあり、LinuxシステムレベルでのSMBリソースの統合が可能です。Sambaクライアントに対していずれのデーモンも実行する必要はありません。

共有

SMBサーバは、そのクライアントに対し、「共有」によってリソースを提供します。共有は、サーバ上のディレクトリ(サブディレクトリを含む)とプリンタです。共有は、「共有名」によってエクスポートされ、この名前でアクセスできます。共有名にはどのように

な名前も設定できます。エクスポートディレクトリの名前である必要はありません。共有プリンタにも名前が割り当てられています。クライアントは名前で共有ディレクトリとプリンタにアクセスできます。

慣例により、ドル記号(\$)で終わる共有名は非表示になります。Windowsコンピュータを使用して、使用可能な共有を参照する際には表示されません。

DC

ドメインコントローラ(DC)は、ドメインのアカウントを処理するサーバです。データレプリケーションの場合、單一ドメインに複数のドメインコントローラを含めることができます。

20.2 Sambaサーバのインストール

Sambaサーバをインストールするには、YaSTを起動して、ソフトウェア › ソフトウェア管理の順に選択します。表示 › パターンの順に選択し、ファイルサーバを選択します。必要なパッケージのインストールを確認して、インストールプロセスを完了します。

20.3 Sambaの起動および停止

Sambaサーバは、自動(ブート中)か手動で起動または停止できます。ポリシーの開始および停止は、[20.4.1項 「YaSTによるSambaサーバの設定」](#)で説明しているように、YaST Sambaサーバ設定の一部です。

コマンドラインから、Sambaに必要なサービスを`systemctl stop smb nmb`を使用して停止し、`systemctl start nmb smb`を使用して開始します。`smb`サービスは、必要に応じて`winbind`を処理します。



ヒント: winbind

`winbind`は、独立したサービスで、個別の`samba-winbind`パッケージとしても提供されます。

20.4 Sambaサーバの設定

SUSE® Linux Enterprise ServerのSambaサーバは、YaSTを使って、または手動で設定することができます。手動で設定を行えば細かい点まで調整できますが、YaSTのGUIほど便利ではありません。

20.4.1 YaSTによるSambaサーバの設定

Sambaサーバを設定するには、YaSTを起動して、ネットワークサービス › Sambaサーバの順に選択します。

20.4.1.1 初期Samba設定

このモジュールを初めて起動すると、Sambaインストールダイアログが起動して、サーバ管理に関するいくつかの基本的な事項を決定するように要求されます。設定の最後に、Samba管理者パスワードを要求されます(Sambaルートパスワード)。次回起動時には、Samba Configurationダイアログが表示されます。

Sambaインストールダイアログは、次の2つのステップとオプションの詳細設定で構成されています。

ワークグループまたはドメイン名

Workgroup or Domain Nameから既存の名前を選択するか、新しい名前を入力し、次へを入力します。

Sambaサーバのタイプ

次のステップでは、サーバをPDC(プライマリドメインコントローラ)として機能させるか、BDC(バックアップドメインコントローラ)として機能させるか、またはドメインコントローラとしては機能させないかを指定します。次へで続行します。

詳細なサーバ設定に進まない場合は、OKを選択して確認します。次に、最後のポップアップボックスで、Sambaルートパスワードを設定します。

この設定はすべて、後からSambaの設定ダイアログで起動、共有、識別情報、信頼されたドメイン、LDAP設定の各タブを使用して変更することができます。

20.4.1.2 サーバ上でSMBプロトコルの現在のバージョンを有効にする

現在のバージョンのSUSE Linux Enterprise Serverまたは他の最新のLinuxバージョンを実行しているクライアントで、安全ではないSMB1/CIFSプロトコルはデフォルトで無効になっています。ただし、Sambaの既存のインスタンスはSMB1/CIFSバージョンのプロトコルを使用する共有にのみサービスを提供するように設定できます。このようなクライアントとやり取りするためには、少なくともSMB 2.1プロトコルを使用して共有にサービスを提供するようにSambaを設定する必要があります。

たとえば、SMB1/CIFSのUnix拡張機能に依存する、SMB1のみが使用可能な設定があります。これらの拡張機能は、より新しいバージョンのプロトコルには移植されていません。このような状況にある場合は、設定を変更することを検討するか、[20.5.2項 「クライアント上へのSMB1/CIFS共有のマウント」](#)を参照してください。

これを行うには、設定ファイル/etc/samba/smb.confで、グローバルパラメータserver max protocol = SMB2_10を設定します。すべての可能な値のリストについては、[man smb.conf](#)を参照してください。

20.4.1.3 Sambaの詳細設定

Sambaサーバモジュールの初回起動中、2つの初期化ステップ([20.4.1.1項 「初期Samba設定」](#)参照)の直後にSambaの設定ダイアログが表示されます。ここでは、Sambaサーバの設定を編集することができます。

設定を編集し終わったら、OKをクリックして設定を保存します。

20.4.1.3.1 サーバを起動する

Start Upタブで、Sambaサーバの起動に関する設定を行います。システムのブート時に毎回サービスが起動されるようにするには、During Bootを選択します。手動起動を有効化するには、Manuallyを選択します。Sambaサーバの起動の詳細については、[20.3項 「Sambaの起動および停止」](#)を参照してください。

このタブで、ファイアウォールのポートを開くこともできます。そのためには、Open Port in Firewallを選択します。複数のネットワークインターフェースがある場合は、Firewall Detailsをクリックし、インターフェースを選択した後、OKをクリックして、Sambaサービス用のネットワークインターフェースを選択します。

20.4.1.3.2 共有

共有タブで、有効にするSambaの共有を指定します。homesおよびプリンタなど、事前定義済みの共有がいくつかあります。状態の変更を使用して、有効と無効の間で切り替えます。新規の共有を追加するには追加、共有を削除するには削除をクリックします。

ユーザにディレクトリの共有を許可するを選択すると、許可するグループ中のグループメンバーに、各自のディレクトリを他のユーザと共有させることができます。たとえば、ローカルの範囲のusers、あるいはドメインの範囲ではDOMAIN\Usersを設定します。また、ユーザ

にはファイルシステムへのアクセスを許可するパーミッションがあることを確認してください。最大共有数で、共有の最大数を制限することができます。認証なしでユーザ共用へのアクセスを許可するには、ゲストアクセスを許可を有効にします。

20.4.1.3.3 ID

識別情報タブで、ホストが関連付けられているドメイン(基本設定)と、ネットワークで代替ホスト名を使用するかどうか(NetBIOSホスト名)を指定します。名前解決にMicrosoft Windows Internet Name Service(WINS)を使用することもできます。この場合、Use WINS for Hostname Resolutionを有効にし、DHCP経由でWINSサーバを取得(Retrieve WINS server via DHCPを使用)するかどうか決定します。TDBデータベースではなくLDAPなど、エキスパートグローバル設定またはユーザ認証ソースを設定するには、詳細設定をクリックします。

20.4.1.3.4 信頼されたドメイン

他のドメインのユーザを、自分のドメインにアクセスさせるには、Trusted Domainsタブで適切な設定を行います。追加をクリックして、新しいドメインを追加します。選択したドメインを削除するには、削除をクリックします。

20.4.1.3.5 LDAP設定

LDAP Settingsタブでは、認証に使用するLDAPサーバを設定することができます。LDAPサーバへの接続をテストするには、Test Connectionをクリックします。エキスパートLDAP設定を設定するか、デフォルト値を使用する場合、詳細な設定をクリックします。

LDAP設定に関する詳細については、『Security and Hardening Guide』、第5章「LDAP with 389 Directory Server」を参照してください。

20.4.2 サーバの手動設定

Sambaをサーバとして使用する場合は、sambaをインストールします。Sambaの主要設定ファイルは、/etc/samba/smb.confです。このファイルは2つの論理部分に分けられます。[global]セクションには、中心的なグローバル設定が含まれます。次のデフォルトのセクションには、個別のファイルとプリンタ共有が入っています。

- [homes]
- [プロファイル]

- [users]
- [グループ]
- [プリンタ]
- [印刷\$]

この方法を使用すると、共有のオプションを [global] セクションで別々にまたはグローバルに設定することができます。これにより、環境設定ファイルが理解しやすくなります。

20.4.2.1 グローバルセクション

[global] セクションの次のパラメータは、ネットワークの設定に応じた必要条件を満たし、Windows環境で他のマシンがSMBを経由してこのSambaサーバにアクセスできるようにするために変更が必要です。

workgroup = WORKGROUP

この行は、Sambaサーバをワークグループに割り当てます。WORKGROUPを実際のネットワーク環境にある適切なワークグループに置き換えてください。DNS名がネットワーク内の他のマシンに割り当てられていなければ、SambaサーバがDNS名の下に表示されます。DNS名が使用できない場合は、netbiosname=MYNAMEを使用してサーバ名を設定します。このパラメータに関する詳細については、smb.confのマニュアルページを参照してください。

os level = 20

このパラメータは、SambaサーバがワークグループのLMB(ローカルマスタブラウザ)になるかどうかのきっかけとなります。Sambaサーバの設定が誤っていた場合に、既存のWindowsネットワークに支障が出ないよう、小さな値(たとえば2)を選択します。このトピックの詳細については、『Samba 3 Howto』の「ネットワークブラウジング」の章を参照してください。『Samba 3 Howto』の詳細については、20.9項 「詳細情報」 を参照してください。

ネットワーク内に他のSMBサーバ(たとえば、Windows 2000サーバ)が存在せず、ローカル環境に存在するすべてのシステムのリストをSambaサーバに保存する場合は、os levelの値を大きくします(たとえば、65)。これでSambaサーバが、ローカルネットワークのLMBとして選択されました。

この設定を変更するときは、それが既存のWindowsネットワーク環境にどう影響するかを慎重に検討する必要があります。はじめに、隔離されたネットワークで、または影響の少ない時間帯に、変更をテストしてください。

wins support および wins server

アクティブなWINSサーバをもつ既存のWindowsネットワークにSambaサーバを参加させる場合は、wins serverオプションを有効にし、その値をWINSサーバのIPアドレスに設定します。

各Windowsマシンの接続先サブネットが異なり、互いを認識させなければならぬ場合は、WINSサーバをセットアップする必要があります。SambaサーバをWINSサーバなどにするには、wins support = Yesオプションを設定します。ネットワーク内でこの設定が有効なSambaサーバは1台だけであることを確認します。オプションwins serverとwins supportは、smb.confファイル内で同時に有効にしないでください。

20.4.2.2 共有

次の例では、SMBクライアントがCD-ROMドライブとユーザディレクトリ(homes)を利用できるようにする方法を示します。

[cdrom]

CD-ROMドライブが誤って利用可能になるのを避けるため、これらの行はコメントマーク(この場合はセミコロン)で無効にします。最初の列のセミコロンを削除し、CD-ROMドライブをSambaと共有します。

例 20.1: CD-ROMの共有

```
[cdrom]
comment = Linux CD-ROM
path = /media/cdrom
locking = No
```

[cdrom] および comment

[cdrom]セクションエントリは、ネットワーク上のすべてのSMBクライアントが認識できる共有の名前です。さらにcommentを追加して、共有を説明することができます。

path = /media/cdrom

pathは、/media/cdromディレクトリをエクスポートします。

デフォルトを非常に制約的に設定することによって、このシステム上に存在するユーザのみがこの種の共有を利用できるようになります。この共有をあらゆるユーザに開放する場合は、設定にguest ok = yesという行を追加します。この設定は、ネットワーク上の全ユーザに読み込み許可を与えます。このパラメータは慎重に扱うことをお勧めします。またこのパラメータを[u global]セクションで使用する場合には、さらに注意が必要です。

[homes]

[homes]共有は、ここでは特に重要です。ユーザがLinuxファイルサーバの有効なアカウントとパスワードを持ち、独自のホームディレクトリを持っていればそれに接続することができます。

例 20.2: [HOMES]共有

```
[homes]
comment = Home Directories
valid users = %S
browseable = No
read only = No
inherit acls = Yes
```

[homes]

SMBサーバに接続しているユーザの共有名を他の共有が使用していない限り、[homes]共有ディレクティブを使用して共有が動的に生成されます。生成される共有の名前は、ユーザ名になります。

valid users = %S

%Sは、接続が正常に確立されたときに、具体的な共有名に置き換えられます。[homes]共有の場合、これは常にユーザ名です。したがって、ユーザの共有に対するアクセス権は、そのユーザだけに付与されます。

browseable = No

この設定を行うと、共有がネットワーク環境で認識されなくなります。

read only = No

デフォルトでは、Sambaはread only = Yesパラメータによって、エクスポートされた共有への書き込みアクセスを禁止します。共有に書き込めるように設定するには、read only = No値を設定します。これはwritable = Yesと同値です。

create mask = 0640

MS Windows NTベース以外のシステムは、UNIXのパーミッションの概念を理解しないので、ファイルの作成時にパーミッションを割り当てることができません。create maskパラメータは、新しく作成されたファイルに割り当てられるアクセス権を定義します。これは書き込み可能な共有にのみ適用されます。実際には、この設定は、所有者に読み込み/書き込み権限があり、所有者のプライマリグループのメンバーに読み込み権限があることを意味しています。valid users = %Sは、グループに読み込み権限がある場合でも読み込みアクセスを禁止します。グループに読み書き権を付与する場合は、valid users = %Sという行を無効にしてください。



警告: NFSマウントをSambaと共有しない

NFSマウントのSambaとの共有は、データが失われる可能性があるため、サポートされていません。ファイルサーバにSambaを直接インストールするか、iSCSIなどの代替方法を使用することを検討してください。

20.4.2.3 セキュリティレベル

セキュリティを向上させるため、各共有へのアクセスは、パスワードによって保護されています。SMBでは、次の方法で権限を確認できます。

ユーザレベルセキュリティ (security = user)

このセキュリティレベルは、ユーザという概念をSMBに取り入れています。各ユーザは、サーバにパスワードを登録する必要があります。登録後、エクスポートされた個々の共有へのアクセスは、ユーザ名に応じてサーバが許可します。

ADSレベルセキュリティ (security = ADS)

このモードでは、Sambaはアクティブディレクトリ環境のドメインメンバーとして動作します。このモードで操作するには、Sambaを実行しているコンピュータにKerberosがインストールされ設定済みであることが必要です。Sambaを使用してコンピュータをADSレベルムに結合させる必要があります。これは、YaSTのWindowsドメインメンバー・シップモジュールを使用して行います。

ドメインレベルセキュリティ (security = domain)

このモードは、マシンがWindows NTドメインに参加している場合にのみ正しく動作します。Sambaは、Windows Serverと同じ方法で、ユーザ名とパスワードをWindowsプライマリまたはバックアップドメインコントローラに渡すことで検証しようとします。暗号化されたパスワードパラメータがyesに設定されている必要があります。

共有、ユーザ、サーバ、またはドメインレベルのセキュリティの設定は、サーバ全体に適用されます。個別の共有ごとに、ある共有には共有レベルのセキュリティ、別の共有にはユーザレベルセキュリティを設定するといったことはできません。しかし、システム上に設定したIPアドレスごとに、別のSambaサーバを実行することは可能です。

この詳細については、『Samba 3 HOWTO』を参照してください。1つのシステムで複数のサーバを使用する場合は、オプションinterfacesとbind interfaces onlyに注意してください。

20.5 クライアントの設定

クライアントは、TCP/IP経由でのみSambaサーバにアクセスできます。IPX経由のNetBEUIおよびNetBIOSは、Sambaで使用できません。

20.5.1 YaSTによるSambaクライアントの設定

SambaクライアントをSambaサーバまたはWindowsサーバ上のリソース(ファイルまたはプリント)にアクセスするように設定します。WindowsまたはActive Directoryのドメインまたはワークグループを、ネットワークサービス › Windowsドメインメンバーシップの順に選択して表示したダイアログに入力します。Linuxの認証にもSMBの情報を使用するを有効にした場合、ユーザ認証は、Samba、Windows、またはKerberosのサーバ上で実行されます。

エキスパート設定をクリックして、高度な設定オプションを設定します。たとえば、認証による自動的なサーバホームディレクトリのマウントを有効化するには、サーバディレクトリのマウントのテーブルを使用します。これにより、CIFS上でホストされると、ホームディレクトリにアクセスできるようになります。詳細については、[pam_mount](#)のマニュアルページを参照してください。

すべての設定を完了したら、ダイアログを確認して設定を終了します。

20.5.2 クライアント上へのSMB1/CIFS共有のマウント

SMBネットワークプロトコルの最初のバージョン、SMB1またはCIFSは、古くて安全ではないプロトコルであるため、開発者であるMicrosoftによって推奨されていません。セキュリティ上の理由から、SUSE Linux Enterprise Serverの[mount](#)コマンドは、デフォルトでより新しいプロトコルバージョン(SMB 2.1、SMB 3.0、またはSMB 3.02)のみを使用して、SMB共有をマウントします。

ただし、この変更は、[mount](#)、および[/etc/fstab](#)を介したマウントにのみ影響します。SMB1は、明示的に要求することで引き続き使用できます。使用する情報は、以下のとおりです。

- [smbclient](#)ツール。
- SUSE Linux Enterprise Serverに付属するSambaサーバソフトウェア。

SMB1のみ使用可能なため、このデフォルト設定により接続障害が生じる次のような設定があります。

- より新しいSMBプロトコルバージョンをサポートしないSMBサーバを使用した設定。Windowsでは、Windows 7およびWindows Server 2008以降、SMB 2.1のサポートを提供しています。
- SMB1/CIFSのUnix拡張機能に依存する設定。これらの拡張機能は、より新しいバージョンのプロトコルには移植されていません。

! 重要: システムセキュリティの低減

下に記載される指示に従うと、セキュリティの問題に対処できる場合があります。問題に関する詳細については、<https://blogs.technet.microsoft.com/filecab/2016/09/16/stop-using-smb1/> を参照してください。

できるだけ早くサーバをアップグレードすると、より安全なSMBバージョンにすることができます。

SUSE Linux Enterprise Serverで適切なプロトコルバージョンを有効化する方法については、20.4.1.2項 「サーバ上でSMBプロトコルの現在のバージョンを有効にする」を参照してください。

現在のSUSE Linux Enterprise ServerカーネルでSMB1共有を有効にする必要がある場合は、使用するmountコマンドラインにvers=1.0オプションを追加します。

```
# mount -t cifs //HOST/Sshare /MOUNT_POINT -o username=USER_ID,vers=1.0
```

または、SUSE Linux Enterprise Serverのインストール内でSMB1共有をグローバルに有効にすることもできます。有効にするには、/etc/samba/smb.confの[u_global]セクションの下に次のコマンドを追加します。

```
client min protocol = CORE
```

20.6 口グインサーバとしてのSamba

ビジネス設定では、セントラルインスタンスで登録されているユーザにのみアクセスを許可するのが望ましい場合が多いです。Windowsベースのネットワークでは、このタスクはPDC(プライマリドメインコントローラ)によって処理されます。WindowsサーバをPDCとして使用することができますが、Sambaサーバを使用しても処理できます。例20.3 「smb.confファイルのグローバルセクション」に示すように、smb.confの[u_global]セクションにエントリを追加する必要があります。

例 20.3: SMB.CONFファイルのグローバルセクション

```
[global]
```

```
workgroup = WORKGROUP
domain logons = Yes
domain master = Yes
```

ユーザーアカウントとパスワードをWindowsに準拠した暗号化形式で作成する必要があります。そのためにはコマンド smbpasswd -a name を実行します。さらに次のコマンドを使用して、Windowsドメイン概念で必要になるコンピュータのドメインアカウントを作成します。

```
useradd hostname
smbpasswd -a -m hostname
```

useradd コマンドを使用すると、ドル記号が追加されます。コマンド smbpasswd を指定すると、パラメータ -m を使用したときにドル記号が自動的に挿入されます。コメント付きの設定例 (/usr/share/doc/packages/samba/examples/smb.conf.SUSE) には、この作業を自動化するための設定が含まれています。

```
add machine script = /usr/sbin/useradd -g nogroup -c "NT Machine Account" \
-s /bin/false %m
```

Sambaがこのスクリプトを正常に実行できるようにするために、必要な管理者権限を持つ Sambaユーザを選択して、ntadmin グループに追加します。これにより、このLinuxグループに属するすべてのユーザに対し、次のコマンドによって Domain Admin ステータスを割り当てることができます。

```
net groupmap add ntgroup="Domain Admins" unixgroup=ntadmin
```

20.7 Active Directoryネットワーク内のSambaサーバ

LinuxサーバとWindowsサーバの両方を利用する場合、2つの独立した認証システムまたはネットワークを作成するか、または単一の中央認証システムを持つ単一のネットワークに両方のサーバを接続します。SambaはActive Directoryドメイン(AD)と連携できるため、お使いのSUSE Linux Enterprise ServerをActive Directoryドメインに参加させることができます。

Active Directoryドメインに参加させるには、次の手順に従います。

1. root としてログインし、YaSTを起動します。
2. ネットワークサービス › Windowsドメインメンバーシップの順に選択します。
3. Windowsドメインメンバーシップ画面のドメインまたはワークグループフィールドに、参加するドメインを入力します。



図 20.1: WINDOWS ドメインメンバーシップの決定

4. ServerでLinux認証にSMBソースを使用する場合は、Linuxの認証にもSMBの情報を用いるを選択します。
5. ドメインへの参加を確認するメッセージが表示されたら、OKをクリックします。
6. Active DirectoryサーバのWindows管理者用パスワードを入力し、OKをクリックします。
Active Directoryドメインコントローラから、すべての認証データを取得できるようになりました。

または、realmdツールを使用して、Active Directoryに接続することもできます。詳細については、20.7.1項 「realmdを使用したActive Directoryの管理」 を参照してください。



ヒント: 識別情報マッピング

複数のSambaサーバが存在する環境では、UIDとGIDが常に作成されるわけではありません。ユーザに割り当てるUIDは、最初のログイン順になるため、サーバ間でUIDの競合が生じます。この問題を解決するには、識別情報マッピングを利用する必要があります。詳しくは<https://www.samba.org/samba/docs/man/Samba-HOWTO-Collection/idmapper.html>を参照してください。

20.7.1 **realmd**を使用したActive Directoryの管理

realmdは、ネットワーク認証とドメインメンバーシップを設定できるDBusサービスです。

20.7.1.1 Active Directory ドメインの検出

realmdは、DNS SRVレコードをチェックして、使用または設定可能なドメインまたはレルムを検出します。検出するActive Directoryドメイン(次の例では**domain.example.com**)に対してDNS SRVレコードが使用可能であることを確認します。

```
_ldap._tcp.dc._msdcs.domain.example.com.
```

DNSレコードは、Active Directory付属のDNSサーバによって自動的に作成されるはずです。

特定のドメイン名を検出するには、次のコマンドを実行します。

```
> sudo realm discover --verbose domain.example.com

* Resolving: _ldap._tcp.dc._msdcs.domain.example.com
* Sending MS-CLDAP ping to: 192.168.20.10
* Sending MS-CLDAP ping to: 192.168.12.12
* Successfully discovered: domain.example.com
...
```

特定のActive Directory ドメインに参加させるには、次のコマンドを実行します。

```
> sudo realm join --verbose domain.example.com
```

Active Directory ドメインに参加させたら、ドメインアカウントを使用してログインできるようにマシンを設定できます。このためには、次のコマンドを実行します。

```
> sudo realm permit --realm domain.example.com --all
```

特定のアカウントのみを許可するには、次のようにコマンドでそれを指定します。

```
> sudo realm permit --realm domain.example.com DOMAIN\\USERNAME DOMAIN\\USERNAME
```

任意のドメインアカウントからのログインを拒否するには、次のコマンドを使用します。

```
> sudo realm deny --realm domain.example.com --all
```

20.8 詳細トピック

このセクションでは、Sambaスイートのクライアントとサーバの両方の部分を管理するためのより高度なテクニックを紹介します。

20.8.1 systemdを使用したCIFSファイルシステムの自動化

systemdを使用して起動時にCIFS共有をマウントできます。そのためには、以下の説明に従つて進めます。

1. マウントポイントを作成します。

```
> mkdir -p PATH_SERVER_SHARED_FOLDER
```

ここで、PATH_SERVER_SHARED_FOLDERは、以降のステップでは/cifs/sharedです。

2. systemdユニットファイルを作成し、前のステップで指定したパスからファイル名を生成します。ここで、「/」は「-」に置き換えられます。例:

```
> sudo touch /etc/systemd/system/cifs-shared.mount
```

次の内容が続きます。

```
[Unit]
Description=CIFS share from The-Server

[Mount]
What=//The-Server/Shared-Folder
Where=/cifs/shared
Type=cifs
Options=rw,username=vagrant,password=admin

[Install]
WantedBy=multi-user.target
```

3. サービスを有効化します。

```
> sudo systemctl enable cifs-shared.mount
```

4. サービスを開始します。

```
> sudo systemctl start cifs-shared.mount
```

サービスが実行中であることを確認するには、次のコマンドを実行します。

```
> sudo systemctl status cifs-shared.mount
```

5. CIFS共有パスが使用可能であることを確認するには、次のコマンドを実行します。

```
> cd /cifs/shared
> ls -l
```

```
total 0
-rwxrwxrwx. 1 root      root      0 Oct 24 22:31 hello-world-cifs.txt
drwxrwxrwx. 2 root      root      0 Oct 24 22:31 subfolder
-rw-r--r--. 1 vagrant   vagrant   0 Oct 28 21:51 testfile.txt
```

20.8.2 Btrfsでの透過的なファイル圧縮

Sambaでは、クライアントは、Btrfsファイルシステムに配置されている共有のファイルおよびディレクトリの圧縮フラグをリモートで操作できます。Windowsエクスプローラでは、ファイル > プロパティ > 詳細ダイアログを使用することで、ファイル/ディレクトリに透過的な圧縮対象のフラグを付けることができます。

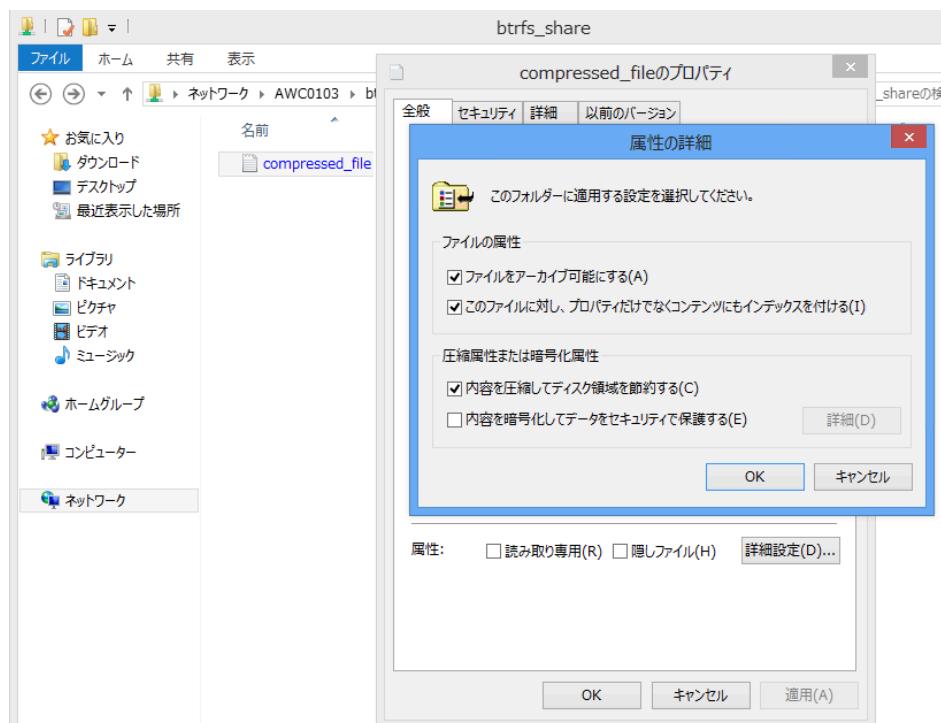


図 20.2: WINDOWSエクスプローラの属性の詳細ダイアログ

圧縮対象フラグが付いたファイルは、アクセスまたは変更があると、基礎となるファイルシステムによって透過的に圧縮および圧縮解除されます。通常、これによってファイルアクセス時に余分なCPUオーバーヘッドが生じますが、ストレージ容量の節約になります。新しいファイルとディレクトリは、FILE_NO_COMPRESSIONオプションを指定して作成しない限り、親ディレクトリの圧縮フラグを継承します。

Windowsエクスプローラでは、圧縮ファイルとディレクトリは、未圧縮のファイル/ディレクトリとは視覚的に見分けが付くように表示されます。

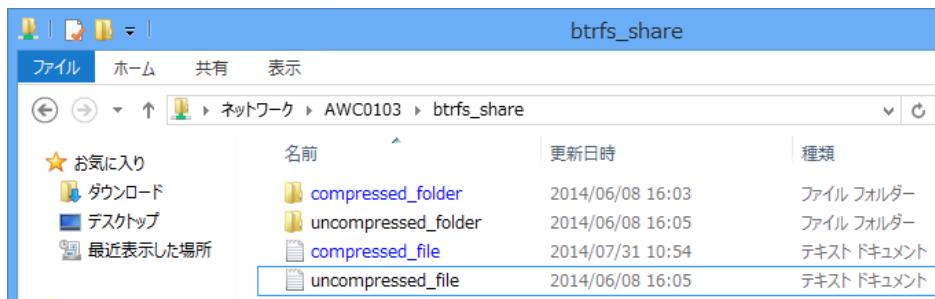


図 20.3: WINDOWSエクスプローラでの圧縮ファイルのディレクトリリスト

Samba共有の圧縮を有効にするには、手動で、

```
vfs objects = btrfs
```

/etc/samba/smb.confに共有設定を追加して実行するか、YaSTを使用してネットワークサービス › Sambaサーバ › 追加の順に選択してbtrfsの機能を使用するをオンにします。

Btrfsでの圧縮の概要については、[1.2.2.1項 「圧縮されたBtrfsファイルシステムのマウント」](#)を参照してください。

20.8.3 スナップショット

スナップショット(シャドウコピーとも呼ばれる)は、特定の時点におけるファイルシステムサブボリュームの状態のコピーです。Snapperは、Linuxでこれらのスナップショットを管理するためのツールです。スナップショットは、BtrfsファイルシステムまたはシンプロビジョンングされたLVMボリュームでサポートされています。Sambaスイートは、サーバ側とクライアント側の両方で、FSRVPプロトコルを介したリモートスナップショットの管理をサポートしています。

20.8.3.1 以前のバージョン

Sambaサーバ上のスナップショットは、以前のバージョンのファイルまたはディレクトリとしてリモートWindowsクライアントに公開できます。

Sambaサーバでスナップショットを有効にするには、次の条件を満たしている必要があります。

- SMBネットワーク共有がBtrfsサブボリューム上に存在している。
- SMBネットワーク共有のパスに、関連するSnapper環境設定ファイルが含まれている。
次のコマンドを使用して、Snapperファイルを作成できます。

```
> sudo snapper -c <cfg_name> create-config /path/to/share
```

Snapperの詳細については、『管理ガイド』、第10章「Snapperを使用したシステムの回復とスナップショット管理」を参照してください。

- スナップショットディレクトリツリーでは、関連するユーザにアクセスを許可する必要があります。詳細については、マニュアルページの「PERMISSIONS」のセクション([man 8 vfs_snapper](#))を参照してください。

リモートスナップショットをサポートするには、[/etc/samba/smb.conf](#)ファイルを変更する必要があります。変更するには、YaST › ネットワークサービス › Sambaサーバの順に選択するか、または次のコマンドを使用して関連する共有セクションを手動で拡張します。

```
vfs objects = snapper
```

手動での[smb.conf](#)への変更を有効にするために、Sambaサービスを再起動する必要がある点に注意してください。

```
> sudo systemctl restart nmb smb
```



図 20.4: **スナップショットが有効な新しいSAMBA共有の追加**

設定後、Samba共有パスでSnapperによって作成されたスナップショットには、Windowsエクスプローラのファイルまたはディレクトリの以前のバージョンタブからアクセスできます。

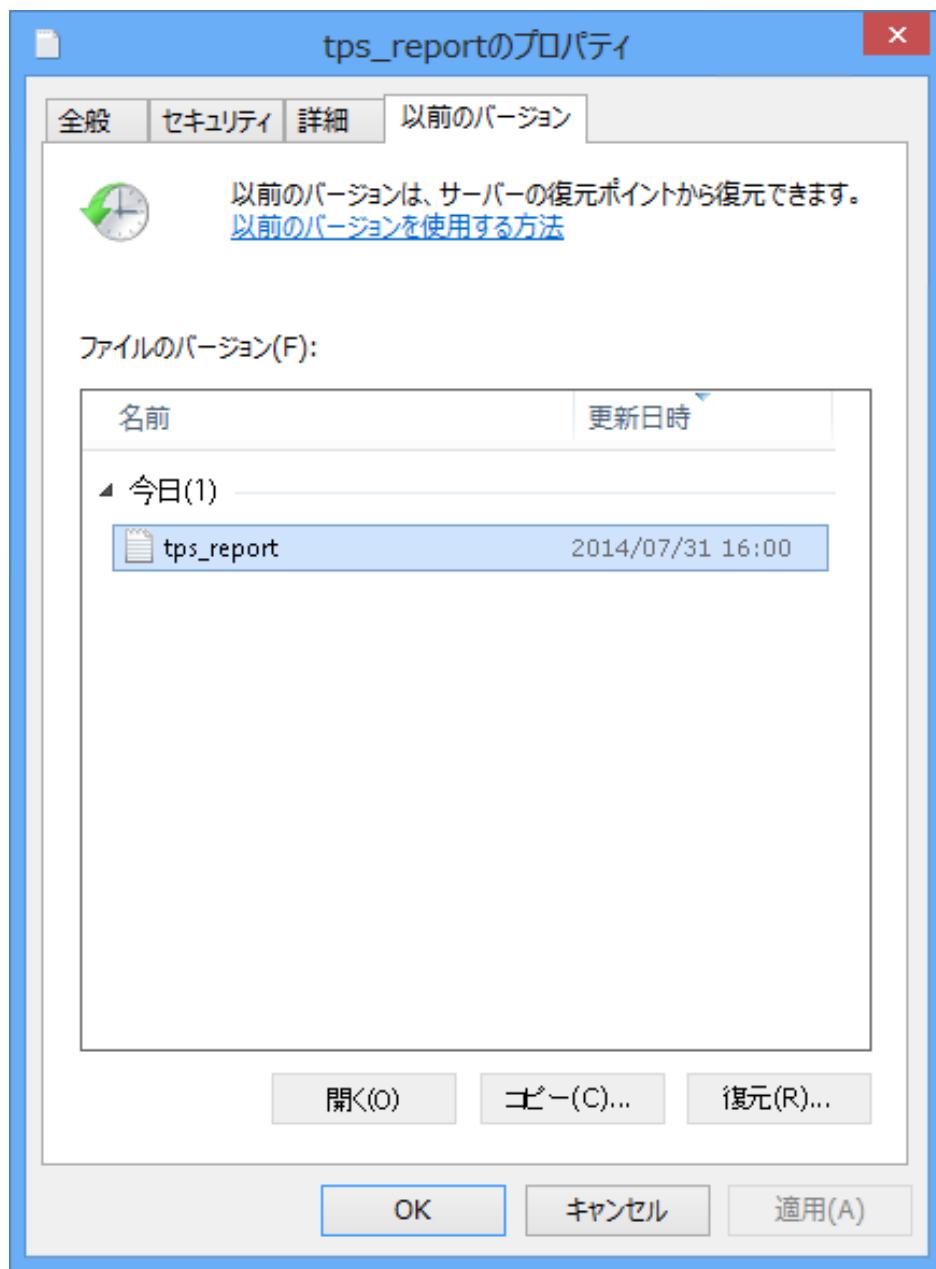


図 20.5: WINDOWSエクスプローラの以前のバージョンタブ

20.8.3.2 リモート共有スナップショット

デフォルトでは、スナップショットは、SnapperコマンドラインユーティリティまたはSnapperのタイムライン機能を使用して、Sambaサーバ上でローカルでのみ作成および削除できます。

Sambaは、リモートホストからの共有スナップショット作成および削除要求をFSRVP (File Server Remote VSS Protocol)を使用して処理するように設定できます。

20.8.3.1項「以前のバージョン」で説明されている環境設定と前提条件に加え、/etc/samba/smb.confに次のグローバル設定が必要です。

```
[global]
rpc_daemon:fssd = fork
registry shares = yes
include = registry
```

その後、FSRVPクライアント(SambaのrpcclientおよびWindows Server 2012 DiskShadow.exeを含む)は、特定の共有のスナップショットを作成または削除したり、スナップショットを新しい共有として公開したりするようSambaに命令できます。

20.8.3.3 rpcclientによるLinuxからのスナップショットのリモート管理

samba-clientパッケージには、特定の共有のスナップショットの作成と公開をWindows/Sambaサーバにリモートで要求できるFSRVPクライアントが含まれています。SUSE Linux Enterprise Serverの既存のツールを使用して、公開された共有をマウントし、そのファイルをバックアップできます。サーバへの要求は、rpcclientバイナリを使用して送信されます。

例 20.4: rpcclientを使用したWINDOWS SERVER 2012共有スナップショットの要求

win-server.example.comサーバにEXAMPLEドメインの管理者として接続します。

```
# rpcclient -U 'EXAMPLE\Administrator' ncacn_np:win-server.example.com[ndr64,sign]
Enter EXAMPLE/Administrator's password:
```

rpcclientにSMB共有が表示されることを確認します。

```
# rpcclient $> netshareenum
netname: windows_server_2012_share
remark:
path:   C:\Shares\windows_server_2012_share
password:      (null)
```

SMB共有がスナップショットの作成をサポートしていることを確認します。

```
# rpcclient $> fss_is_path_sup windows_server_2012_share \
UNC \\WIN-SERVER\windows_server_2012_share\ supports shadow copy requests
```

共有スナップショットの作成を要求します。

```
# rpcclient $> fss_create_expose backup ro windows_server_2012_share
13fe880e-e232-493d-87e9-402f21019fb6: shadow-copy set created
13fe880e-e232-493d-87e9-402f21019fb6(1c26544e-8251-445f-be89-d1e0a3938777): \
\\WIN-SERVER\windows_server_2012_share\ shadow-copy added to set
13fe880e-e232-493d-87e9-402f21019fb6: prepare completed in 0 secs
13fe880e-e232-493d-87e9-402f21019fb6: commit completed in 1 secs
```

```
13fe880e-e232-493d-87e9-402f21019fb6(1c26544e-8251-445f-be89-d1e0a3938777): \
share windows_server_2012_share@{1C26544E-8251-445F-BE89-D1E0A3938777} \
exposed as a snapshot of \\WIN-SERVER\windows_server_2012_share\
```

スナップショット共有がサーバによって公開されたことを確認します。

```
# rpcclient $> netshareenum
netname: windows_server_2012_share
remark:
path:   C:\Shares\windows_server_2012_share
password:      (null)

netname: windows_server_2012_share@{1C26544E-8251-445F-BE89-D1E0A3938777}
remark: (null)
path:  \\?\GLOBALROOT\Device\HarddiskVolumeShadowCopy{F6E6507E-F537-11E3-9404-
B8AC6F927453}\Shares\windows_server_2012_share\
password:      (null)
```

スナップショット共有の削除を試みます。

```
# rpcclient $> fss_delete windows_server_2012_share \
13fe880e-e232-493d-87e9-402f21019fb6 1c26544e-8251-445f-be89-d1e0a3938777
13fe880e-e232-493d-87e9-402f21019fb6(1c26544e-8251-445f-be89-d1e0a3938777): \
\\WIN-SERVER\windows_server_2012_share\ shadow-copy deleted
```

スナップショット共有がサーバによって削除されたことを確認します。

```
# rpcclient $> netshareenum
netname: windows_server_2012_share
remark:
path:   C:\Shares\windows_server_2012_share
password:      (null)
```

20.8.3.4 **DiskShadow.exe**によるWindowsからのスナップショットのリモート管理

WindowsクライアントからLinux Samba上でSMB共有のスナップショットを管理することができます。Windows Server 2012には、[20.8.3.3項 「rpcclientによるLinuxからのスナップショットのリモート管理」](#)で説明した**rpcclient**コマンドと同様にリモート共有を管理できる**DiskShadow.exe**ユーティリティが含まれています。最初にSambaサーバを慎重に設定する必要がある点に注意してください。

以下は、Windowsクライアントが共有のスナップショットを管理できるようにSambaサーバを設定する手順の例です。EXAMPLEはテスト環境で使用されるActive Directoryドメイン、fsrvp-server.example.comはSambaサーバのホスト名、/srv/smbはSMB共有のパスである点に注意してください。

手順 20.1: SAMBAサーバの詳細な設定

1. YaSTを介してActive Directoryドメインに参加します。詳細については、[20.7項「Active Directoryネットワーク内のSambaサーバ」](#)を参照してください。
2. Active DirectoryドメインのDNSエントリが正しいことを確認します。

```
fsrvp-server:~ # net -U 'Administrator' ads dns register \
fsrvp-server.example.com <IP address>
Successfully registered hostname with DNS
```

3. Btrfsサブボリュームを/srv/smbに作成します。

```
fsrvp-server:~ # btrfs subvolume create /srv/smb
```

4. パス/srv/smbにSnapper環境設定ファイルを作成します。

```
fsrvp-server:~ # snapper -c <snapper_config> create-config /srv/smb
```

5. パス/srv/smbで新しい共有を作成し、YaSTのスナップショットの公開チェックボックスをオンにします。[20.8.3.2項「リモート共有スナップショット」](#)に説明されているように、次のスニペットを/etc/samba/smb.confのグローバルセクションに追加します。

```
[global]
rpc_daemon:fssd = fork
registry_shares = yes
include = registry
```

6. systemctl restart nmb smbでSambaを再起動します。

7. Snapperのパーミッションを設定します。

```
fsrvp-server:~ # snapper -c <snapper_config> set-config \
ALLOW_USERS="EXAMPLE\\Administrator EXAMPLE\\win-client$"
```

ALLOW_USERSのすべてのインスタンスが.snapshotsサブディレクトリへのアクセスも許可されていることを確認します。

```
fsrvp-server:~ # snapper -c <snapper_config> set-config SYNC_ACL=yes
```



重要: パスのエスケープ

「\」エスケープには注意してください。/etc/snapper/configs/<snapper_config>に保存された値を確実に1回エスケープするには、2回エスケープします。

「EXAMPLE\win-client\$」はWindowsクライアントのコンピュータアカウントに対応します。Windowsは、このアカウントが認証されている間に初期FSRVP要求を発行します。

8. Windowsクライアントアカウントに必要な特権を付与します。

```
fsrvp-server:~ # net -U 'Administrator' rpc rights grant \
"EXAMPLE\win-client$" SeBackupPrivilege
Successfully granted rights.
```

「EXAMPLE\Administrator」ユーザの場合、すでに特権が付与されているため、上のコマンドは必要ありません。

手順 20.2: **WINDOWSクライアントのセットアップとDiskShadow.exeの実行**

1. Windows Server 2012 (ホスト名の例:WIN-CLIENT)をブートします。
2. SUSE Linux Enterprise Serverと同じActive DirectoryドメインEXAMPLEに参加します。
3. 再起動します。
4. PowerShellを開きます。
5. DiskShadow.exeを起動し、バックアップ手順を開始します。

```
PS C:\Users\Administrator.EXAMPLE> diskshadow.exe
Microsoft DiskShadow version 1.0
Copyright (C) 2012 Microsoft Corporation
On computer: WIN-CLIENT, 6/17/2014 3:53:54 PM

DISKSHADOW> begin backup
```

6. シャドウコピーがプログラムの終了、リセット、および再起動にわたって維持されるように指定します。

```
DISKSHADOW> set context PERSISTENT
```

7. 指定した共有がスナップショットをサポートしているかどうかを確認し、スナップショットを作成します。

```
DISKSHADOW> add volume \\fsrvp-server\sles_snapper

DISKSHADOW> create
Alias VSS_SHADOW_1 for shadow ID {de4ddca4-4978-4805-8776-cdf82d190a4a} set as \
environment variable.
Alias VSS_SHADOW_SET for shadow set ID {c58e1452-c554-400e-a266-d11d5c837cb1} \
set as environment variable.
```

```
Querying all shadow copies with the shadow copy set ID \
{c58e1452-c554-400e-a266-d11d5c837cb1}

* Shadow copy ID = {de4ddca4-4978-4805-8776-cdf82d190a4a}      %VSS_SHADOW_1%
- Shadow copy set: {c58e1452-c554-400e-a266-d11d5c837cb1}  %VSS_SHADOW_SET%
- Original count of shadow copies = 1
- Original volume name: \\FSRVP-SERVER\SLES_SNAPPER\ \
[volume not on this machine]
- Creation time: 6/17/2014 3:54:43 PM
- Shadow copy device name:
\\FSRVP-SERVER\SLES_SNAPPER@{31af84a-44a7-41be-b9b0-751898756faa}
- Originating machine: FSRVP-SERVER
- Service machine: win-client.example.com
- Not exposed
- Provider ID: {89300202-3cec-4981-9171-19f59559e0f2}
- Attributes: No_Auto_Release Persistent FileShare

Number of shadow copies listed: 1
```

8. バックアップ手順を終了します。

```
DISKSHADOW> end backup
```

9. スナップショットが作成された後、その削除を試み、削除されたことを確認します。

```
DISKSHADOW> delete shadows volume \\FSRVP-SERVER\SLES_SNAPPER\
Deleting shadow copy {de4ddca4-4978-4805-8776-cdf82d190a4a} on volume \
\\FSRVP-SERVER\SLES_SNAPPER\ from provider \
{89300202-3cec-4981-9171-19f59559e0f2} [Attributes: 0x04000009]...

Number of shadow copies deleted: 1

DISKSHADOW> list shadows all

Querying all shadow copies on the computer ...
No shadow copies found in system.
```

20.9 詳細情報

- **マニュアルページ:** `man`パッケージでインストールされるすべての`samba`ページのリストを表示するには、`apropos samba`を実行します。`man NAME_OF_MAN_PAGE`を使用してマニュアルページを開きます。
- **SUSE-specific READMEファイル:** パッケージ`samba-client`には`/usr/share/doc/packages/samba/README.SUSE`ファイルが含まれています。

- **追加のパッケージドキュメント:** `zypper install samba-doc`でパッケージ `samba-doc` をインストールします。
このマニュアルは `/usr/share/doc/packages/samba` にインストールされます。マニュアルページのHTMLバージョンと設定例のライブラリ (`smb.conf.SUSE` など) が含まれています。
- **オンラインマニュアル:** Samba wikiには、広範囲なUser Documentation (https://wiki.samba.org/index.php/User_Documentation) が含まれています。

21 autofsによるオンデマンドマウント

autofsは、指定したディレクトリをオンデマンドベースで自動的にマウントするプログラムです。これは高い効率を実現するためにカーネルモジュールに基づいており、ローカルディレクトリとネットワーク共有の両方を管理できます。これらの自動的なマウントポイントは、アクセスがあった場合にのみマウントされ、非アクティブな状態が一定時間続くとアンマウントされます。このオンデマンドの動作によって帯域幅が節約され、/etc/fstabで管理する静的マウントよりも高いパフォーマンスが得られます。autofsは制御スクリプトですが、automountは実際の自動マウントを実行するコマンド(デーモン)です。

21.1 インストール

デフォルトでは、autofsはSUSE Linux Enterprise Serverにインストールされません。その自動マウント機能を利用するには、最初に、次のコマンドを使用してインストールします。

```
> sudo zypper install autofs
```

21.2 設定

vimなどのテキストエディタで設定ファイルを編集して、autofsを手動で設定する必要があります。autofsの基本的な設定手順は2つあります。「マスタ」マップファイルを使用する手順と、特定のマップファイルを使用する手順です。

21.2.1 マスタマップファイル

autofsのデフォルトのマスタ設定ファイルは/etc/auto.masterです。その場所を変更するには、/etc/sysconfig/autofs内のDEFAULT_MASTER_MAP_NAMEオプションの値を変更します。次に、SUSE Linux Enterprise Serverのデフォルトのマスタ設定ファイルの内容を示します。

```
#  
# Sample auto.master file  
# This is an automounter map and it has the following format  
# key [ -mount-options-separated-by-comma ] location
```

```

# For details of the format look at autofs(5). ①
#
#/misc /etc/auto.misc ②
#/net -hosts
#
# Include /etc/auto.master.d/*.autofs ③
#
#+dir:/etc/auto.master.d
#
# Include central master map if it can be found using
# nsswitch sources.
#
# Note that if there are entries for /net or /misc (as
# above) in the included master map any keys that are the
# same will not be seen as the first read key seen takes
# precedence.
#
+auto.master ④

```

- ① 自動マウント機能のマップの形式については、[autofs](#)のマニュアルページ([man 5 autofs](#))で多くの貴重な情報が提供されています。
- ② デフォルトではコメント化(#)されていますが、これは単純な自動マウント機能のマッピング構文の例です。
- ③ マスタマップファイルを複数のファイルに分割する必要がある場合、この行のコメント化を解除し、マッピング(サフィックスは[.autofs](#))を[/etc/auto.master.d/](#)ディレクトリに配置します。
- ④ [+auto.master](#)により、NIS (NISの詳細については、『Security and Hardening Guide』、第3章「Using NIS」、3.1項「Configuring NIS servers」を参照)を使用してもそのマスタマップが確実に見つかるようになります。

[auto.master](#)のエントリには3つのフィールドがあり、構文は次のとおりです。

mount point	map name	options
-------------	----------	---------

mount point

[autofs](#)ファイルシステムをマウントする基本の場所(/[home](#)など)。

map name

マウントに使用するマップソースの名前。マップファイルの構文については、[21.2.2項「マップファイル」](#)を参照してください。

options

これらのオプションを指定した場合、指定したマップ内のすべてのエントリにデフォルトとして適用されます。



ヒント: 詳細情報

オプションの `map-type`、`format`、および `options` の特定の値の詳細について
は、`auto.master` のマニュアルページ ([man 5 auto.master](#)) を参照してください。

`auto.master` の次のエントリは、`autofs` に対し、`/etc/auto.smb` 内を検索して `/smb` ディレクトリにマウントポイントを作成するよう指示します。

```
/smb    /etc/auto.smb
```

21.2.1.1 直接マウント

直接マウントは、関連するマップファイル内で指定されたパスにマウントポイントを作成します。`auto.master` でマウントポイントを指定するのではなく、マウントポイントフィールドを `/-` に置き換えます。たとえば、次の行は、`autofs` に対し、`auto.smb` で指定された場所にマウントポイントを作成するよう指示します。

```
/-      /etc/auto.smb
```



ヒント: フルパスを使用しないマップ

ローカルまたはネットワークのフルパスでマップファイルを指定していない場合、マップファイルはネームサービススイッチ (NSS) 設定を使用して検索されます。

```
/-      auto.smb
```

21.2.2 マップファイル



重要: 他のタイプのマップ

`autofs` による自動マウントのマップタイプとしては「ファイル」が最も一般的ですが、他のタイプもあります。マップは、コマンドの出力や、LDAP またはデータベースのクエリ結果で指定することもできます。マップタイプの詳細については、[man 5 auto.master](#) マニュアルページを参照してください。

マップファイルは、ソースの場所(ローカルまたはネットワーク)と、ソースをローカルにマウントするためのマウントポイントを指定します。マップの全般的な形式はマスタマップと同様です。異なるのは、**options**をエントリの最後ではなくmount pointとlocationの間に記述する点です。

mount point	options	location
-------------	---------	----------

マップファイルが実行可能ファイルとしてマークされていないことを確認してください。`chmod -x MAP_FILE`を実行することにより、実行可能ビットを削除することができます。

mount point

ソースの場所をどこにマウントするかを指定します。ここには、[で指定されたベースマウントポイントに追加する1つのディレクトリ名\(auto.master 「間接」マウント\)](#)、またはマウントポイントのフルパス(直接マウント、[21.2.1.1項 「直接マウント」を参照](#))のいずれかを指定できます。

options

関連するエントリのマウントオプションを、カンマで区切ったオプションのリストで指定します。このマップファイルのオプションも[auto.master](#)に含まれている場合、これらが追加されます。

location

ファイルシステムのマウント元の場所を指定します。通常は、標準の表記方法[host_name:path_name](#)によるNFSまたはSMBボリュームです。マウントするファイルシステムが「/」で始まる場合(ローカルの[/dev](#)エントリやsmbfs共有など)、[:/dev/sda1](#)のように、コロン記号「:」のプレフィックスを付ける必要があります。

21.3 操作とデバッグ

このセクションでは、[autofs](#)サービスの操作を制御する方法と、自動マウント機能の操作を調整する際に詳細なデバッグ情報を表示する方法の概要について説明します。

21.3.1 autofsサービスの制御

[autofs](#)サービスの動作は、[systemd](#)によって制御されます。[autofs](#)用の[systemctl](#)コマンドの一般的な構文は、次のとおりです。

> <code>sudo systemctl SUB_COMMAND autofs</code>
--

ここでSUB_COMMANDは以下のいずれかです。

enable

ブート時に自動マウント機能のデーモンを起動します。

start

自動マウント機能のデーモンを起動します。

stop

自動マウント機能のデーモンを停止します。自動マウントポイントにはアクセスできません。

status

autofsサービスの現在のステータスと、関連するログファイルの一部を出力します。

restart

自動マウント機能を停止して起動します。実行中のデーモンをすべて終了し、新しいデーモンを起動します。

reload

現在のauto.masterマップを確認して、エントリに変更があるデーモンを再起動し、新しいエントリがある場合は新しいデーモンを起動します。

21.3.2 自動マウント機能の問題のデバッグ

autofsでディレクトリをマウントする際に問題が発生する場合は、automountデーモンを手動で実行して出力メッセージを確認してください。

1. autofsを停止します。

```
> sudo systemctl stop autofs
```

2. 1つの端末から、フォアグラウンドでautomountを手動で実行し、詳細な出力を生成します。

```
> sudo automount -f -v
```

3. 別の端末から、マウントポイントにアクセスして(たとえば、cdまたはlsを使用して)、自動マウントファイルシステムをマウントしてみます。

4. 1番目の端末から、automountの出力で、マウントに失敗した理由またはマウントが試行されていない理由についての詳細情報がないかどうかを確認します。

21.4 NFS共有の自動マウント

次の手順は、ネットワーク上で利用可能なNFS共有を自動マウントするよう`autofs`を設定する方法を示しています。この方法は上で説明した情報を利用しています。また、NFSのエクスポートを熟知していることが前提です。NFSの詳細については、[第19章 「NFS共有ファイルシステム」](#)を参照してください。

1. マスタマップファイル`/etc/auto.master`を編集します。

```
> sudo vim /etc/auto.master
```

`/etc/auto.master`の最後に新しいNFSマウント用の新しいエントリを追加します。

```
/nfs      /etc/auto.nfs      --timeout=10
```

これは、ベースマウントポイントは`/nfs`で、NFS共有は`/etc/auto.nfs`マップで指定されていることを`autofs`に伝え、非アクティブな状態が10秒間続いたらこのマップ内のすべての共有を自動的にアンマウントするよう指示します。

2. NFS共有用の新しいマップファイルを作成します。

```
> sudo vim /etc/auto.nfs
```

通常、`/etc/auto.nfs`には、各NFS共有に対して別個の行が含まれます。形式については、[21.2.2項 「マップファイル」](#)を参照してください。マウントポイントおよびNFS共有のネットワークアドレスを記述する行を追加します。

```
export    jupiter.com:/home/geeko/doc/export
```

上述の行は、要求があると、`jupiter.com`ホスト上の`/home/geeko/doc/export`ディレクトリがローカルホスト上の`/nfs/export`ディレクトリ(`/nfs`は`auto.master`マップから取得)に自動マウントされることを意味します。`/nfs/export`ディレクトリは、`autofs`によって自動的に作成されます。

3. 以前に同じNFS共有を静的にマウントしていた場合、必要に応じて`/etc/fstab`の関連する行をコメント化します。行は次のようにになります。

```
#jupiter.com:/home/geeko/doc/export /nfs/export nfs defaults 0 0
```

4. `autofs`を再ロードし、動作しているかどうかを確認します。

```
> sudo systemctl restart autofs
```

```
# ls -l /nfs/export
```

```
total 20
drwxr-xr-x  5 1001 users 4096 Jan 14  2017 .images/
drwxr-xr-x 10 1001 users 4096 Aug 16  2017 .profiled/
drwxr-xr-x  3 1001 users 4096 Aug 30  2017 .tmp/
drwxr-xr-x  4 1001 users 4096 Apr 25 08:56 manual/
```

リモート共有上にあるファイルのリストを参照できる場合、autofsは機能しています。

21.5 詳細トピック

このセクションでは、autofsの基本的な説明よりも詳しいトピックについて説明します。ここで説明するのは、ネットワーク上で利用可能なNFS共有の自動マウント、マップファイルでのワイルドカードの使用、およびCIFSファイルシステムに固有の情報です。

21.5.1 /net mount point

このヘルパーマウントポイントは、大量のNFS共有を使用する場合に便利です。/netには、ローカルネットワーク上にあるすべてのNFS共有がオンデマンドで自動マウントされます。このエントリはすでにauto.masterファイルに存在しているため、エントリのコメント化を解除してautofsを再起動するだけで済みます。

```
/net      -hosts
> sudo systemctl restart autofs
```

たとえば、jupiterという名前のサーバと/exportという名前のNFS共有がある場合、

```
> sudo cd /net/jupiter/export
```

コマンドラインで次のように入力してマウントできます。

21.5.2 ワイルドカードを使用したサブディレクトリの自動マウント

個別に自動マウントする必要があるサブディレクトリが含まれるディレクトリがある場合(代表的なケースは、個々のユーザのホームディレクトリが内部にある/homeディレクトリ)、autofsには便利な解決方法が備わっています。

ホームディレクトリの場合は、auto.masterに次の行を追加します。

```
/home      /etc/auto.home
```

続いて、/etc/auto.homeファイルに正しいマッピングを追加し、ユーザのホームディレクトリが自動的にマウントされるようにする必要があります。1つの解決方法は、各ディレクトリに対して個別のエントリを作成することです。

```
wilber    jupiter.com:/home/wilber
penguin   jupiter.com:/home/penguin
tux       jupiter.com:/home/tux
[...]
```

これは、auto.home内にあるユーザのリストを管理する必要があるため、効率的とはいえません。マウントポイントの代わりにアスタリスク「*」を使用し、マウントするディレクトリの代わりにアンパサンド「&」を使用します。

```
*      jupiter:/home/&
```

21.5.3 CIFSファイルシステムの自動マウント

SMB/CIFS共有を自動マウントする場合(SMB/CIFSプロトコルの詳細については、[第20章「Samba」](#)を参照)、マップファイルの構文を変更する必要があります。オプションフィールドに-fstype=cifsを追加し、共有の場所にコロン「:」のプレフィックスを付けます。

```
mount point      -fstype=cifs      ://jupiter.com/export
```

A GNU licenses

改訂履歷

2023-02-03

This appendix contains the GNU Free Documentation License version 1.2.

GNU Free Documentation License

Copyright (C) 2000, 2001, 2002 Free Software Foundation, Inc. 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA. Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or non-commercially. Secondly, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software. We have designed this License to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or non-commercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other

material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- H. Include an unaltered copy of this License.
- I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

- L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
- N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
- O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties--for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what

the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See <https://www.gnu.org/copyleft/>.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

```
Copyright (c) YEAR YOUR NAME.  
Permission is granted to copy, distribute and/or modify this document  
under the terms of the GNU Free Documentation License, Version 1.2  
or any later version published by the Free Software Foundation;  
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.  
A copy of the license is included in the section entitled "GNU  
Free Documentation License".
```

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

```
with the Invariant Sections being LIST THEIR TITLES, with the  
Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.
```

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.