



SUSE Linux Enterprise High Availability Extension
12 SP5

管理ガイド


管理ガイド

SUSE Linux Enterprise High Availability Extension 12 SP5

著者: Tanja Roth, Thomas Schraitle


このガイドは、SUSE® Linux Enterprise High Availability Extensionを使用してクラスタを設定、構成、および管理する必要がある管理者を対象にしています。構成と管理をすばやく効率的に行うため、High Availability Extensionにはグラフィカルユーザインタフェース(GUI)とコマンドラインインタフェース(CLI)の両方が備わっています。主要なタスクである、両方のアプローチ(GUIおよびCLI)の実行については、このガイドで詳細に説明されています。これにより、管理者は、ニーズを満たす適切なツールを選択できるようになります。

発行日: 2025 年 4 月 03 日

SUSE LLC
1800 South Novell Place
Provo, UT 84606
USA
<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、<http://www.suse.com/company/legal/> を参照してください。その他の製品名および会社名は、各社の商標または登録商標です。商標記号(®、™など)は、SUSEおよび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

このガイドについて xv

- 1 利用可能なマニュアル xvi
- 2 フィードバック xvii
- 3 マニュアルの表記規則 xviii
- 4 本マニュアルの作成について xix

I インストール、セットアップ、およびアップグレード 1

1 製品の概要 2

- 1.1 拡張としての提供 2
- 1.2 主な機能 2
 - 広範なクラスタリングシナリオ 3 • 柔軟性 3 • ストレージとデータレプリケーション 3 • 仮想化環境のサポート 4 • ローカル、メトロ、およびGeoクラスタのサポート 4 • リソースエージェント 5 • ユーザフレンドリな管理ツール 5
- 1.3 利点 6
- 1.4 クラスタ設定: ストレージ 9
- 1.5 アーキテクチャ 12
 - アーキテクチャ層 12 • プロセスフロー 14

2 システム要件と推奨事項 16

- 2.1 ハードウェア要件 16
- 2.2 ソフトウェアの必要条件 17
- 2.3 ストレージ要件 17
- 2.4 その他の要件と推奨事項 18

3	High Availability Extensionのインストール	20
3.1	手動インストール	20
3.2	AutoYaSTによる大量インストールと展開	20
4	YaSTクラスタモジュールの使用	22
4.1	用語の定義	22
4.2	YaSTクラスタモジュール	24
4.3	通信チャンネルの定義	26
4.4	認証設定の定義	30
4.5	すべてのノードへの設定の転送	31
	YaSTによるCsync2の設定	32
	• Csync2を使用した変更内容の同期	33
4.6	クラスタノード間の接続ステータスの同期	35
4.7	サービスの設定	36
4.8	クラスタをオンラインにする	38
5	クラスタアップグレードとソフトウェアパッケージの更新	39
5.1	用語集	39
5.2	最新の製品バージョンへのクラスタアップグレード	40
	SLE HAおよびSLE HA Geoでサポートされるアップグレードパス	41
	• アップグレード前に必要な準備	44
	• オフラインマイグレーション	44
	• ローリングアップグレード	47
5.3	クラスタノード上のソフトウェアパッケージの更新	48
5.4	その他の情報	49
II	設定および管理	50
6	設定および管理の基本事項	51
6.1	ユースケースのシナリオ	51

- 6.2 クォーラムの判断 52
 - グローバルクラスタオプション 53 • グローバルオプションno-quorum-policy 53 • グローバルオプションstonith-enabled 54 • 2ノードクラスタのCorosync設定 54 • NノードクラスタのCorosync設定 55
- 6.3 クラスタリソース 56
 - リソース管理 56 • サポートされるリソースエージェントクラス 57 • リソースのタイプ 58 • リソーステンプレート 59 • 高度なリソースタイプ 60 • リソースオプション(メタ属性) 63 • インスタンス属性(パラメータ) 66 • リソース操作 68 • タイムアウト値 69
- 6.4 リソース監視 71
- 6.5 リソースの制約 72
 - 制約のタイプ 72 • スコアと無限大 75 • リソーステンプレートと制約 76 • フェールオーバーノード 76 • フェールバックノード 78 • 負荷インパクトに基づくリソースの配置 78 • タグの使用によるリソースのグループ化 81
- 6.6 リモートホストでのサービスの管理 82
 - 監視プラグインを使用したリモートホストでのサービスの監視 82 • pacemaker_remoteを使用したリモートノードでのサービスの管理 84
- 6.7 システムヘルスの監視 84
- 6.8 その他の情報 86
- 7 Hawk2を使用したクラスタリソースの設定と管理 88
 - 7.1 Hawk2の要件 88
 - 7.2 ログイン 89
 - 7.3 Hawk2の概要: 主な構成要素 90
 - 左のナビゲーションバー 90 • 最上位の行 91
 - 7.4 グローバルクラスタオプションの設定 92
 - 7.5 クラスタリソースの設定 94
 - 現在のクラスタ設定の表示(CIB) 95 • ウィザードを使用したリソースの追加 95 • 単純なリソースの追加 96 • リソーステンプレートの追加 98 • リソースの変更 98 • STONITHリソースの追加 100 • ク

- ラスタリソースグループの追加 101 • クローンリソースの追加 103 • マルチステートリソースの追加 104 • タグの使用によるリソースのグループ化 105 • リソース監視の設定 106
- 7.6 制約の設定 108
 - 場所制約の追加 109 • コロケーション制約の追加 110 • 順序制約の追加 112 • 制約のためにリソースセットを使用する 114 • その他の情報 115 • リソースフェールオーバーノードの指定 116 • リソースフェールバックノードの指定(リソースの固着性) 117 • 負荷インパクトに基づくリソース配置の設定 118
- 7.7 クラスタリソースの管理 120
 - リソースとグループの編集 120 • リソースの開始 121 • リソースのクリーンアップ 122 • クラスタリソースの削除 122 • クラスタリソースの移行 123
- 7.8 クラスタの監視 124
 - 単一クラスタの監視 124 • 複数のクラスタの監視 126
- 7.9 バッチモードの使用 129
- 7.10 クラスタ履歴の表示 133
 - ノードまたリソースの最近のイベントの表示 133 • クラスタレポートのための履歴エクスプローラーの使用 134 • 履歴エクスプローラーの遷移詳細の表示 136
- 7.11 クラスタヘルスの確認 137
- 8 クラスタリソースの設定と管理(コマンドライン) 139
 - 8.1 crmsh - 概要 139
 - ヘルプの表示 140 • crmshのサブコマンドの実行 141 • OCFリソースエージェントに関する情報の表示 142 • crmshのシェルスクリプトの使用 144 • crmshのクラスタスクリプトの使用 144 • 設定テンプレートの使用 147 • シャドーイング設定のテスト 149 • 設定の変更のデバッグ 150 • クラスタダイアグラム 150
 - 8.2 Corosync設定の管理 151
 - 8.3 グローバルクラスタオプションの設定 152
 - 8.4 クラスタリソースの設定 152
 - ファイルからのクラスタリソースのロード 153 • クラスタリソースの作成 153 • リソーステンプレートの作成 154 • STONITHリソースの作成 155 • リソース制約の設定 156 • リソースフェールオーバーノードの指

	定 159 • リソースフェールバックノードの指定(リソースの固着性) 159 • 負荷インパクトに基づくリソース配置の設定 159 • リソース監視の設定 162 • クラスタリソースグループの構成 162 • クローンリソースの設定 163
8.5	クラスタリソースの管理 164 クラスタリソースの表示 164 • 新しいクラスタリソースの開始 165 • リソースのクリーンアップ 166 • クラスタリソースの削除 166 • クラスタリソースのマイグレーション 167 • リソースのグループ化/タグ付け 167 • ヘルスステータスの取得 167
8.6	cib.xmlから独立したパスワードの設定 168
8.7	履歴情報の取得 169
8.8	詳細 170
9	リソースエージェントの追加または変更 171
9.1	STONITHエージェント 171
9.2	OCFリソースエージェントの作成 171
9.3	OCF戻りコードと障害回復 172
10	フェンシングとSTONITH 175
10.1	フェンシングのクラス 175
10.2	ノードレベルのフェンシング 176 STONITHデバイス 176 • STONITHの実装 177
10.3	STONITHのリソースと環境設定 178 STONITHリソースの設定例 178
10.4	フェンシングデバイスの監視 181
10.5	特殊なフェンシングデバイス 182
10.6	基本的な推奨事項 184
10.7	詳細 184
11	ストレージ保護とSBD 186
11.1	概念の概要 186

- 11.2 SBDの手動設定の概要 188
- 11.3 要件 188
- 11.4 SBDデバイスの数 189
- 11.5 タイムアウトの計算 189
- 11.6 ウォッチドッグのセットアップ 191
 - ハードウェアウォッチドッグの使用 191
 - ソフトウェアウォッチドッグ(softdog)の使用 192
- 11.7 デバイスでのSBDの設定 193
- 11.8 ディスクレスSBDの設定 198
- 11.9 SBDとフェンシングのテスト 199
- 11.10 ストレージ保護のための追加メカニズム 200
 - sg_persistリソースの設定 201
 - sfexを使用した排他的なストレージアクティブ化の保証 202
- 11.11 その他の情報 204
- 12 アクセス制御リスト 205**
 - 12.1 要件と前提条件 205
 - 12.2 クラスタでのACLの使用の有効化 206
 - 12.3 ACLの基⁹⁶事項 207
 - XPath式によるACLルールの設定 207
 - 短縮によるACLルールの設定 209
 - 12.4 Hawk2によるACLの設定 209
 - 12.5 crmshによるACLの設定 211
- 13 ネットワークデバイスボンディング 213**
 - 13.1 YaSTによるボンディングデバイスの設定 213
 - 13.2 ボンディングスレーブのホットプラグ 216
 - 13.3 その他の情報 218

14 負荷バランス 219

- 14.1 概念の概要 219
- 14.2 Linux仮想サーバによる負荷分散の設定 221
 - Director 221 • ユーザスペースのコントローラとデーモン 221 • パケット転送 222 • スケジューリングアルゴリズム 222 • YaSTによるIP負荷分散の設定 223 • 追加設定 227
- 14.3 HAProxyによる負荷分散の設定 228
- 14.4 その他の情報 231

15 Geoクラスタ(マルチサイトクラスタ) 232

16 保守タスクの実行 233

- 16.1 クラスタノードを切断する意味 233
- 16.2 保守タスクのためのさまざまなオプション 234
- 16.3 保守作業の準備と終了 235
- 16.4 クラスタを保守モードにする 235
- 16.5 ノードを保守モードにする 236
- 16.6 ノードをスタンバイモードにする 237
- 16.7 リソースを保守モードにする 237
- 16.8 リソースを非管理対象モードにする 238
- 16.9 保守モード中のクラスタノードの再起動 239

III ストレージおよびデータレプリケーション 240

17 分散ロックマネージャ(DLM:Distributed Lock Manager) 241

- 17.1 DLM通信のプロトコル 241
- 17.2 DLMクラスタリソースの設定 241

18 OCFS2 244

- 18.1 特長と利点 244
- 18.2 OCFS2のパッケージと管理ユーティリティ 245
- 18.3 OCFS2サービスとSTONITHリソースの設定 246
- 18.4 OCFS2ボリュームの作成 247
- 18.5 OCFS2ボリュームのマウント 249
- 18.6 Hawk2でのOCFS2リソースの設定 250
- 18.7 OCFS2ファイルシステム上でクォータを使用する 252
- 18.8 詳細情報 252

19 GFS2 253

- 19.1 GFS2パッケージおよび管理ユーティリティ 253
- 19.2 GFS2サービスとSTONITHリソースの設定 254
- 19.3 GFS2ボリュームの作成 254
- 19.4 GFS2ボリュームのマウント 256

20 DRBD 258

- 20.1 概念の概要 258
- 20.2 DRBDサービスのインストール 259
- 20.3 DRBDサービスの設定 260
 - 手動によるDRBDの設定 261 • YaSTによるDRBDの設定 263 • DRBDリソースの初期化とフォーマット 265
- 20.4 DRBD 8から DRBD 9への移行 266
- 20.5 スタックされたDRBDデバイスの作成 268
- 20.6 リソースレベルのフェンシングの使用 269
- 20.7 DRBDサービスのテスト 270
- 20.8 DRBDのチューニング 272

- 20.9 DRBDのトラブルシュート 272
 - 環境設定 272 • ホスト名 273 • TCPポート7788 273 • DRBDデバイスが再起動後に破損した 273
- 20.10 詳細情報 274
- 21 Cluster Logical Volume Manager(cLVM) 275
 - 21.1 概念の概要 275
 - 21.2 cLVMの環境設定 275
 - クラスタリソースの作成 276 • シナリオ: Cmirrordの設定 276 • シナリオ - SAN上でiSCSIを使用するcLVM 279 • シナリオ - DRBDを使用するcLVM 283
 - 21.3 有効なLVM2デバイスの明示的な設定 284
 - 21.4 詳細 285
- 22 クラスタマルチデバイス(Cluster MD) 286
 - 22.1 概念の概要 286
 - 22.2 クラスタ化されたMD RAIDデバイスの作成 286
 - 22.3 リソースエージェントの設定 288
 - 22.4 デバイスの追加 288
 - 22.5 一時的に障害が発生したデバイスの再追加 289
 - 22.6 デバイスの削除 289
- 23 Sambaクラスタリング 290
 - 23.1 概念の概要 290
 - 23.2 基本的な設定 291
 - 23.3 Active Directoryドメインへの追加 294
 - 23.4 クラスタ対応Sambaのデバッグとテスト 296
 - 23.5 その他の情報 298

24 Relax-and-Recover (Rear)による障害復旧 299

- 24.1 概念の概要 299
 - 障害復旧プランの作成 299
 - ・ 障害復旧とは 300
 - ・ Rearによる障害復旧 300
 - ・ Rearの要件 300
 - ・ Rearバージョンの更新 301
 - ・ Btrfsに伴う制限事項 301
 - ・ シナリオとバックアップのツール 302
 - ・ 基本手順 302
- 24.2 Rearおよびバックアップソリューションのセットアップ 303
- 24.3 復旧インストールシステムの作成 305
- 24.4 復旧プロセスのテスト 305
- 24.5 障害からの復旧 306
- 24.6 その他の情報 306

IV 付録 307

A トラブルシューティング 308

- A.1 インストールと最初のステップ 308
- A.2 ログ記録 309
- A.3 リソース 310
- A.4 STONITHとフェンシング 312
- A.5 履歴 312
- A.6 Hawk2 314
- A.7 その他 314
- A.8 その他の情報 317

B 命名規則 318

C クラスタ管理ツール(コマンドライン) 319

D rootアクセスなしでのクラスタレポートの実行 321

- D.1 ローカルユーザアカウントの作成 321
- D.2 パスワード不要のSSHアカウントの設定 322

- D.3 `sudo`の設定 324
- D.4 クラスタレポートの生成 326

用語集 327

E GNU Licenses 333

このガイドについて

このガイドは、SUSE® Linux Enterprise High Availability Extensionを使用してクラスタを設定、構成、および管理する必要がある管理者を対象にしています。構成と管理をすばやく効率的に行うため、High Availability Extensionにはグラフィカルユーザインタフェース(GUI)とコマンドラインインタフェース(CLI)の両方が備わっています。主要なタスクである、両方のアプローチ(GUIおよびCLI)の実行については、このガイドで詳細に説明されています。これにより、管理者は、ニーズを満たす適切なツールを選択できるようになります。

このガイドは、次のパートで構成されています。

インストール、セットアップ、およびアップグレード

このパートでは、クラスタのインストールと設定を開始する前に、クラスタの基本とアーキテクチャをよく把握し、主要な機能と利点の概要を理解します。必要なハードウェア/ソフトウェア要件と、以降の手順を実行する前に必要な準備作業について学習します。YaSTを使用してHAクラスタのインストールおよび基本セットアップを実行します。クラスタを最新リリースバージョンにアップグレードする方法、または個々のパッケージを更新する方法について学習します。

設定および管理

Webインタフェース(Hawk2)またはコマンドラインインタフェース(crmsh)を使用して、リソースを追加、設定、および管理します。クラスタ設定への不正アクセスを防止するには、役割を定義して、それらを特定のユーザに割り当てることで細かく制御を行います。負荷分散およびフェンシングの使用法を学習します。独自のリソースエージェントの作成、または既存のエージェントの変更を検討している場合、別の種類のリソースエージェントを作成する方法について背景情報を取得できます。

ストレージおよびデータレプリケーション

SUSE Linux Enterprise High Availability Extensionには、クラスタ対応のファイルシステム(OCFS2とGFS2)、およびcLVM (clustered Logical Volume Manager)が標準装備されています。データのレプリケーションでは、DRBD*を使用します。これにより、High Availabilityサービスのデータをクラスタのアクティブノードからスタンバイノードへミラーリングできます。さらに、クラスタ化したSambaサーバにより、異種混合環境にもHigh Availabilityソリューションが提供されます。

付録

一般的な問題とその解決策の概要が記載されています。クラスタ、リソース、および制約に関して、このマニュアルで使用されている命名規則を示します。HA固有の用語を収録した用語集もあります。

このマニュアルの多くの章に、システム上またはインターネットで利用可能な追加のドキュメントリソースへのリンクが含まれています。

1 利用可能なマニュアル



注記: オンラインヘルプと最新のアップデート

製品に関するマニュアルは、<https://documentation.suse.com> からご利用いただけます。最新のアップデートもご利用いただけるほか、マニュアルをさまざまな形式でブラウズおよびダウンロードすることができます。最新のマニュアルアップデートは通常、英語版で検索できます。

この製品の次のマニュアルを入手できます。

インストールおよびセットアップクイックスタート

このマニュアルでは、`ha-cluster-bootstrap` パッケージで提供されているブートストラップスクリプトを使用して、非常に基本的な2ノードクラスタをセットアップする手順を説明します。仮想IPアドレスをクラスタリソースとして設定する手順や、共有ストレージ上でSBDをフェンシングメカニズムとして使用する手順も記載されています。

管理ガイド

このガイドは、SUSE® Linux Enterprise High Availability Extensionを使用してクラスタを設定、構成、および管理する必要がある管理者を対象にしています。構成と管理をすばやく効率的に行うため、High Availability Extensionにはグラフィカルユーザインタフェース(GUI)とコマンドラインインタフェース(CLI)の両方が備わっています。主要なタスクである、両方のアプローチ(GUIおよびCLI)の実行については、このガイドで詳細に説明されています。これにより、管理者は、ニーズを満たす適切なツールを選択できるようになります。

Geo Clusteringのクイックスタート

Geoクラスタリングを使用すると、それぞれ1つのローカルクラスタを備えた地理的に分散された複数のサイトを運用できます。これらのクラスタ間のフェールオーバーは、より高いレベルのエンティティであるブースクラスタチケットマネージャによって管理されます。

Geo Clusteringガイド

Geoクラスタリングを使用すると、それぞれ1つのローカルクラスタを備えた地理的に分散された複数のサイトを運用できます。これらのクラスタ間のフェールオーバーは、より高いレベルのエンティティであるブースクラスタチケットマネージャによって調整されます。このドキュメントでは、ブースのセットアップオプションおよびパラメータ、Geoクラスタ用のCsync2のセットアップ、クラ

スタリソースを設定する方法、および変更時に他のクラスタサイトに転送する方法について詳しく説明します。また、コマンドラインから、およびHawkを使用してGeoクラスタを管理する方法、および最新の製品バージョンにアップグレードする方法についても説明します。

Highly Available NFS Storage with DRBD and Pacemaker

このドキュメントでは、SUSE Linux Enterprise High Availability Extension 12 SP5の次のコンポーネント: DRBD* (Distributed Replicated Block Device)、LVM (Logical Volume Manager)、およびクラスタリソース管理フレームワークであるPacemakerkを使用し、2ノードクラスタの高可用性NFSストレージを設定する方法について説明します。


Pacemakerリモートクイックスタート

このマニュアルでは、Pacemakerと `pacemaker_remote` によって管理される、リモートノードまたはゲストノードを含むHigh Availabilityクラスタをセットアップする手順を説明します。`pacemaker_remote` の「リモート」という用語は、物理的な距離を意味するのではなく、クラスタの「非メンバーシップ」を意味しています。

2 フィードバック

次のフィードバックチャンネルがあります。

サービスおよびサポート

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、<http://www.suse.com/support/>  を参照してください。

製品コンポーネントのバグを報告するには、<https://scc.suse.com/support/requests>  にアクセスしてログインし、[Create New (新規作成)]をクリックします。

バグレポート

SUSE Bugzillaアカウントをお持ちの場合は、このドキュメントのHTMLバージョンの見出し横にある[バグを報告]リンクをクリックしてください。バグレポートを開くことができるBugzillaに移動します。

メール

この製品のドキュメントについてのフィードバックは、`doc-team@suse.com` 宛のメールでも送信できます。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明記してください。エラーの報告または機能拡張の提案では、問題について簡潔に説明し、対応するセクション番号とページ(またはURL)をお知らせください。

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。

- `tux > command`

`root` ユーザを含む、任意のユーザが実行可能なコマンド。

- `root # command`

`root` 特権で実行する必要があるコマンド。多くの場合、これらのコマンドの頭に `sudo` コマンドを置いて実行することもできます。

- `crm(live)`

対話型 `crm` シェルで実行されるコマンド。詳細については、[第8章「クラスタリソースの設定と管理\(コマンドライン\)」](#)を参照してください。

- `/etc/passwd`: デイ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 `PATH`
- `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
- `packagename`: パッケージの名前
- `Alt`、`Alt - F1`: 使用するキーまたはキーの組み合わせ、キーはキーボード上と同様、大文字で表示される
- [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
- `amd64`、`em64t`、`ipf` > この説明は、`amd64`、`em64t`、および `ipf` の各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。◁
- Dancing Penguins (「Penguins」の章、↑ 他のマニュアル): 他のマニュアルの章への参照です。
- 通知



警告

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

続行する前に知っておくべき重要な情報。



注記

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

ガイドラインや実地的なアドバイスなどの役に立つ情報。

クラスターノードと名前、リソース、およびに制約に関する命名規則の概要については、[付録B 命名規則](#)を参照してください。

4 本マニュアルの作成について

このマニュアルは、[DocBook 5 \(http://www.docbook.org\)](http://www.docbook.org) のサブセットであるSUSEDocで作成されています。XMLソースファイルは `jing` (<https://code.google.com/p/jing-trang/> を参照) によって検証され、`xsltproc` によって処理され、Norman Walshによるスタイルシートのカスタマイズ版を使用してXSL-FOに変換されました。最終的なPDFは、[Apache Software Foundation \(https://xmlgraphics.apache.org/fop\)](https://xmlgraphics.apache.org/fop) のFOPを使用して書式設定されています。このマニュアルの作成に使用したオープンソースツールと環境は、DocBook Authoring and Publishing Suite (DAPS) によって提供されたものです。プロジェクトのホームページは<https://github.com/openSUSE/daps> にあります。

このマニュアルのXMLソースコードについては、<https://github.com/SUSE/doc-sleha> を参照してください。

I インストール、セットアップ、およびアップグレード

- 1 製品の概要 2
- 2 システム要件と推奨事項 16
- 3 High Availability Extensionのインストール 20
- 4 YaSTクラスタモジュールの使用 22
- 5 クラスタアップグレードとソフトウェアパッケージの更新 39

1 製品の概要

SUSE® Linux Enterprise High Availability Extensionは、オープンソースクラスタ化技術の統合スイートで、可用性の高い物理Linuxクラスタと仮想Linuxクラスタを実装し、SPOF (シングルポイント障害)をなくします。データ、アプリケーション、サービスなどの重要なリソースの高度な可用性と管理のしやすさを実現します。その結果、ミッションクリティカルなLinuxワークロードに対してビジネスの継続性維持、データ整合性の保護、予期せぬダウンタイムの削減を行います。

基本的な監視、メッセージング、およびクラスタリソース管理の機能を標準装備し、個々の管理対象クラスタリソースのフェールオーバー、フェールバック、およびマイグレーション(負荷分散)をサポートします。

この章では、High Availability Extensionの主な製品機能と利点を紹介します。ここには、いくつかのクラスタ例が記載されており、クラスタを設定するコンポーネントについて学ぶことができます。最後のセクションでは、アーキテクチャの概要を示し、クラスタ内の個々のアーキテクチャ層とプロセスについて説明します。

High Availabilityクラスタのコンテキストでよく使用される用語については、[用語集](#)を参照してください。

1.1 拡張としての提供

High Availability Extensionは、SUSE Linux Enterprise Server 12 SP5の拡張として入手できます。Geo Clustering for SUSE Linux Enterprise High Availability Extensionという、High Availability Extensionの個別の拡張として、地理的に離れたクラスタ(Geoクラスタ)に対するサポートが提供されています。

1.2 主な機能

SUSE® Linux Enterprise High Availability Extensionでは、ネットワークリソースの可用性を確保し、管理することができます。以降のセクションでは、いくつかの主要機能に焦点を合わせて説明します。

1.2.1 広範なクラスタリングシナリオ

High Availability Extensionは次のシナリオをサポートしています。

- アクティブ/アクティブ設定
- アクティブ/パッシブ設定: N+1、N+M、Nから1、NからM
- ハイブリッド物理仮想クラスタ。仮想サーバを物理サーバとともにクラスタ化できます。これによって、サービスの可用性とリソースの使用状況が向上します。
- ローカルクラスタ
- メトロクラスタ(「ストレッチされた」ローカルクラスタ)
- Geoクラスタ(地理的に離れたクラスタ)は、追加のGeo拡張がサポートされます。[1.2.5項「ローカル、メトロ、およびGeoクラスタのサポート」](#)を参照してください。

クラスタには、最大32のLinuxサーバを含めることができます。`pacemaker_remote`を使用すると、この制限を超えて追加のLinuxサーバを含めるようにクラスタを拡張できます。クラスタ内のどのサーバも、クラスタ内の障害が発生したサーバのリソース(アプリケーション、サービス、IPアドレス、およびファイルシステム)を再起動することができます。

1.2.2 柔軟性

High Availability Extensionには、Corosyncメッセージングおよびメンバーシップ層のほか、Pacemakerクラスタリソースマネージャが標準装備されています。Pacemakerの使用によって、管理者は継続的にリソースのヘルスとステータスを監視し、依存関係を管理し、柔軟に設定できるルールとポリシーに基づいてサービスを自動的に開始および停止できます。High Availability Extensionでは、ユーザの組織に適した特定のアプリケーションおよびハードウェアインフラストラクチャに合わせて、クラスタのカスタマイズが可能です。時間依存設定を使用して、サービスを特定の時刻に修復済みのノードに自動的にフェールバック(マイグレート)させることができます。

1.2.3 ストレージとデータレプリケーション

High Availability Extensionでは必要に応じてサーバストレージを自動的に割り当て、再割り当てすることができます。ファイバチャネルストレージエリアネットワーク(SAN)とネットワーク上のiSCSIストレージをサポートします。共有ディスクシステムもサポートされていますが、必要要件ではありません。SUSE Linux Enterprise High Availability Extensionには、クラスタ対応のファイルシステムとボリュームマネージャ(OCFS2)、cLVM (clustered Logical Volume Manager)も含まれていま

す。データのレプリケーションでは、DRBD*を使用して、High Availabilityサービスのデータをクラスターのアクティブノードからスタンバイノードへミラーリングできます。さらに、SUSE Linux Enterprise High Availability Extensionでは、Sambaクラスタリング技術であるCTDB (Clustered Trivial Database)もサポートしています。

1.2.4 仮想化環境のサポート

SUSE Linux Enterprise High Availability Extensionは、物理Linuxサーバと仮想Linuxサーバ両方のクラスタリングをサポートしています。両タイプのサーバの混合もサポートしています。SUSE Linux Enterprise Server 12 SP5には、XenおよびKVM (カーネルベースの仮想マシン)が付属しています。両方がオープンソース仮想ハイパーバイザーです。仮想ゲストシステム(VMとも呼ばれる)はクラスタによるサービスとして管理できます。

1.2.5 ローカル、メトロ、およびGeoクラスタのサポート

SUSE Linux Enterprise High Availability Extensionは、様々な地理的なシナリオをサポートするように拡張されています。Geo Clustering for SUSE Linux Enterprise High Availability Extensionという、High Availability Extensionの個別の拡張として、地理的に離れたクラスタ(Geoクラスタ)に対するサポートが提供されています。

ローカルクラスタ

1つのロケーション内の単一のクラスタ(たとえば、すべてのノードが1つのデータセンターにある)。クラスタはノード間の通信にマルチキャストまたはユニキャストを使用し、フェールオーバーを内部で管理します。ネットワークの遅延時間は無視できます。ストレージは通常、すべてのノードに同時にアクセスされます。

メトロクラスタ

複数の建物またはデータセンターにわたってストレッチできる単一のクラスタ。クラスタはノード間の通信に通常ユニキャストを使用し、フェールオーバーを内部で管理します。ネットワークの遅延時間は通常は短くなります(約20マイルの距離で<5ms)。ストレージは可能な場合はファイバチャネルで接続されます。データレプリケーションは内部でストレージごとに、またはクラスタの管理下でホストベースのミラーリングごとに実行されます。

Geoクラスタ(マルチサイトクラスタ)

それぞれにローカルクラスタを持つ、複数の地理的に離れたサイト。サイトはIPによって交信します。サイト全体のフェールオーバーはより高いレベルのエンティティによって調整されます。Geoクラスタは限られたネットワーク帯域幅および高レイテンシに対応する必要があります。ストレージは同期的にレプリケートされます。

個々のクラスターノード間の地理的距離が大きいほど、クラスターが提供するサービスの高可用性を妨げる可能性のある要因が多くなります。ネットワークの遅延時間、限られた帯域幅およびストレージへのアクセスが長距離クラスターの課題として残ります。

1.2.6 リソースエージェント

SUSE Linux Enterprise High Availability Extensionには、Apache、IPv4、IPv6、その他多数のリソースを管理するための膨大な数のリソースエージェントが含まれています。またIBM WebSphere Application Serverなどの一般的なサードパーティアプリケーション用のリソースエージェントも含まれています。ご利用の製品に含まれているOpen Cluster Framework (OCF)リソースエージェントの概要は、[8.1.3項「OCFリソースエージェントに関する情報の表示」](#)で説明される `crm ra` コマンドを使用してください。

1.2.7 ユーザフレンドリな管理ツール

High Availability Extensionは、クラスターの基本的なインストールとセットアップのほか、効果的な設定および管理に使用できる強力なツールセットを標準装備しています。

YaST

一般的なシステムインストールおよび管理用グラフィカルユーザインタフェース。『インストールおよびセットアップクイックスタート』で説明されているように、YaSTを使用して、High Availability ExtensionをSUSE Linux Enterprise Server上にインストールします。YaSTでは、クラスターまたは個々のコンポーネントの設定に役立つように、High Availabilityカテゴリ内の次のモジュールも提供しています。

- クラスター: 基本的なクラスターセットアップ。詳細については、[第4章「YaSTクラスターモジュールの使用」](#)を参照してください。
- DRBD: Distributed Replicated Block Deviceの設定。
- IP負荷分散: Linux仮想サーバまたはHAProxyによる負荷分散の設定。詳細については、[第14章「負荷バランス」](#)を参照してください。

HA Web Konsole (Hawk2)

Linux以外のマシンから、Linuxクラスターを管理できるWebベースのユーザインタフェース。このインタフェースは、システムにグラフィカルユーザインタフェースがない場合も理想的なソリューションです。リソースの作成と設定の手順を順を追って支援し、リソースの起動、中止、移行などの管理作業を容易にします。詳細については、[第7章「Hawk2を使用したクラスターリソースの設定と管理」](#)を参照してください。

リソースを設定し、すべての監視または管理作業を実行する、統合されたパワフルなコマンドラインインタフェースです。詳細については、[第8章「クラスタリソースの設定と管理\(コマンドライン\)」](#)を参照してください。

1.3 利点

High Availability Extensionでは最大 32台のLinuxサーバを可用性の高いクラスタ(HAクラスタ)に設定し、クラスタ内の任意のサーバにリソースをダイナミックに切り替えたり、移動することができます。サーバ障害発生時のリソースの自動マイグレーションの設定ができます。また、ハードウェアのトラブルシューティングやワークロードのバランスをとるために、リソースを手動で移動することもできます。

High Availability Extensionは、コモディティコンポーネントによる高可用性を提供しています。アプリケーションと操作をクラスタに統合することによって、運用コストを削減できます。さらにHigh Availability Extensionでは、クラスタ全体を一元管理し、変化するワークロード要件に応じてリソースを調整することもできます(手動でのクラスタの「負荷分散」)。3ノード以上でクラスタを設定すると、複数のノードが「ホットスペア」を共用できて無駄がありません。

その他にも重要な利点として、予測できないサービス停止を削減したり、ソフトウェアおよびハードウェアの保守やアップグレードのための計画的なサービス停止を削減できる点が挙げられます。

次に、クラスタによるメリットについて説明します。

- 可用性の向上
- パフォーマンスの改善
- 運用コストの低減
- スケーラビリティ
- 障害回復
- データの保護
- サーバの集約
- ストレージの集約

共有ディスクサブシステムにRAIDを導入することによって、共有ディスクの耐障害性を強化できます。次のシナリオは、High Availability Extensionの利点を紹介するものです。

クラスタシナリオ例

サーバ3台でクラスタが設定され、それぞれのサーバにWebサーバをインストールしたと仮定します。クラスタ内の各サーバが、2つのWebサイトをホストしています。各Webサイトのすべてのデータ、グラフィックス、Webページコンテンツは、クラスタ内の各サーバに接続された、共有ディスクサブシステムに保存されています。次の図は、このクラスタのセットアップを示しています。

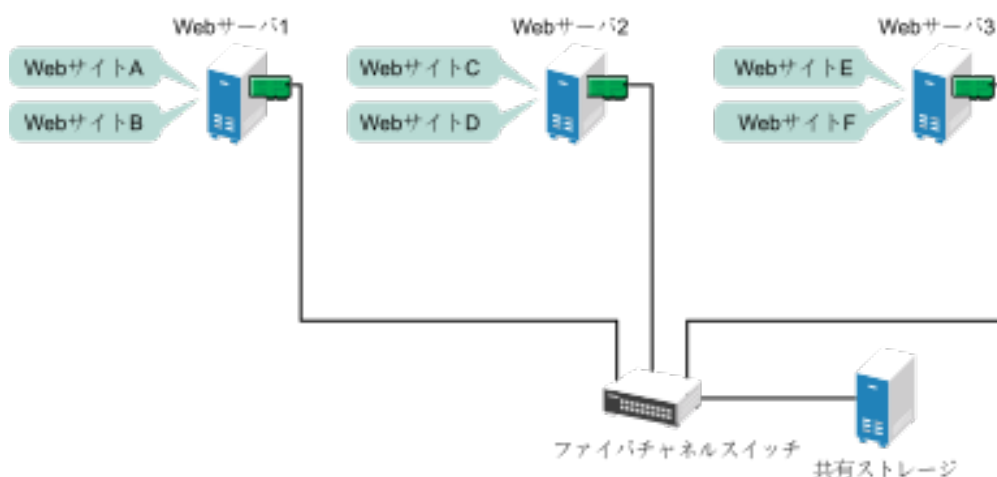


図 1.1: 3サーバクラスタ

通常のクラスタ操作では、クラスタ内の各サーバが他のサーバと常に通信し、すべての登録済みリソースを定期的にポーリングして、障害を検出します。

Webサーバ1でハードウェアまたはソフトウェアの障害が発生したため、このサーバを利用してインターネットアクセス、電子メール、および情報収集を行っているユーザの接続が切断されたとします。次の図は、Webサーバ1で障害が発生した場合のリソースの移動を表したものです。

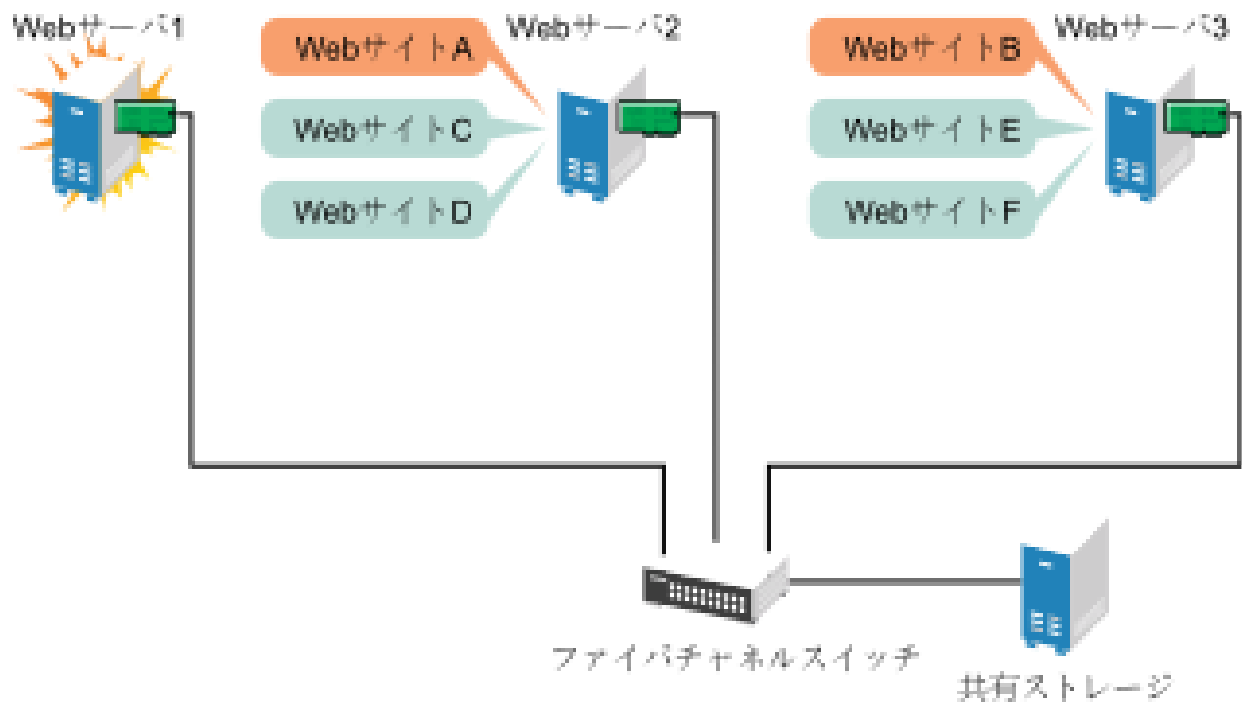


図 1.2: 1台のサーバに障害が発生した後の3サーバクラスタ

WebサイトAがWebサーバ2に、WebサイトBがWebサーバ3に移動します。IPアドレスと証明書もWebサーバ2とWebサーバ3に移動します。

クラスタを設定するときに、それぞれのWebサーバがホストしているWebサイトについて、障害発生時の移動先を指定します。先に説明した例では、WebサイトAの移動先としてWebサーバ2が、WebサイトBの移動先としてWebサーバ3が指定されています。このようにして、Webサーバ1によって処理されていたワークロードが、残りのクラスタメンバーに均等に分散され、可用性を維持できます。

Webサーバ1で障害が発生すると、High Availability Extensionソフトウェアは次の処理を実行します。

- 障害を検出し、Webサーバ1が本当に機能しなくなっていることをSTONITHを使用して検証。STONITHは、「Shoot The Other Node In The Head」(他のノードの頭を撃て)の頭字語であり、誤動作しているノードをダウンさせて、それらがクラスタ内に問題を発生させることを防ぎます。
- Webサーバ1にマウントされていた共有データディレクトリを、Webサーバ2およびWebサーバ3に再マウント。
- Webサーバ1で動作していたアプリケーションを、Webサーバ2およびWebサーバ3で再起動。
- IPアドレスをWebサーバ2およびWebサーバ3に移動。

この例では、フェールオーバープロセスが迅速に実行され、ユーザはWebサイトの情報へのアクセスを数秒程度で回復できます。通常、再度ログインする必要はありません。

ここで、Webサーバ1で発生した問題が解決し、通常に操作できる状態に戻ったと仮定します。WebサイトAおよびWebサイトBは、Webサーバ1に自動的にフェールバック(復帰)することも、そのままの状態を維持することもできます。これは、リソースの設定方法によって決まります。Webサーバ1へのマイグレーションは多少のダウンタイムを伴うため、High Availability Extensionではサービス中断がほとんど、またはまったく発生しないタイミングまでマイグレーションを延期することもできます。いずれの場合でも利点と欠点があります。

High Availability Extensionは、リソースマイグレーション機能も提供します。アプリケーション、Webサイトなどをシステム管理の必要性に応じて、クラスタ内の他のサーバに移動することができます。

たとえば、WebサイトAまたはWebサイトBをWebサーバ1からクラスタ内の他のサーバに手動で移動することができます。これは、Webサーバ1のアップグレードや定期メンテナンスを実施する場合、また、Webサイトのパフォーマンスやアクセスを向上させる場合に有効な機能です。

1.4 クラスタ設定: ストレージ

High Availability Extensionでのクラスタ構成には、共有ディスクサブシステムが含まれる場合と含まれない場合があります。共有ディスクサブシステムの接続には、高速ファイバチャネルカード、ケーブル、およびスイッチを使用でき、また設定にはiSCSIを使用することができます。サーバの障害時には、クラスタ内の別の指定されたサーバが、障害の発生したサーバにマウントされていた共有ディスクディレクトリを自動的にマウントします。この機能によって、ネットワークユーザは、共有ディスクサブシステム上のディレクトリに対するアクセスを中断することなく実行できます。

一般的なリソースの例としては、データ、アプリケーション、およびサービスなどがあります。次の図は、一般的なファイバチャネルクラスタの設定を表したものです。

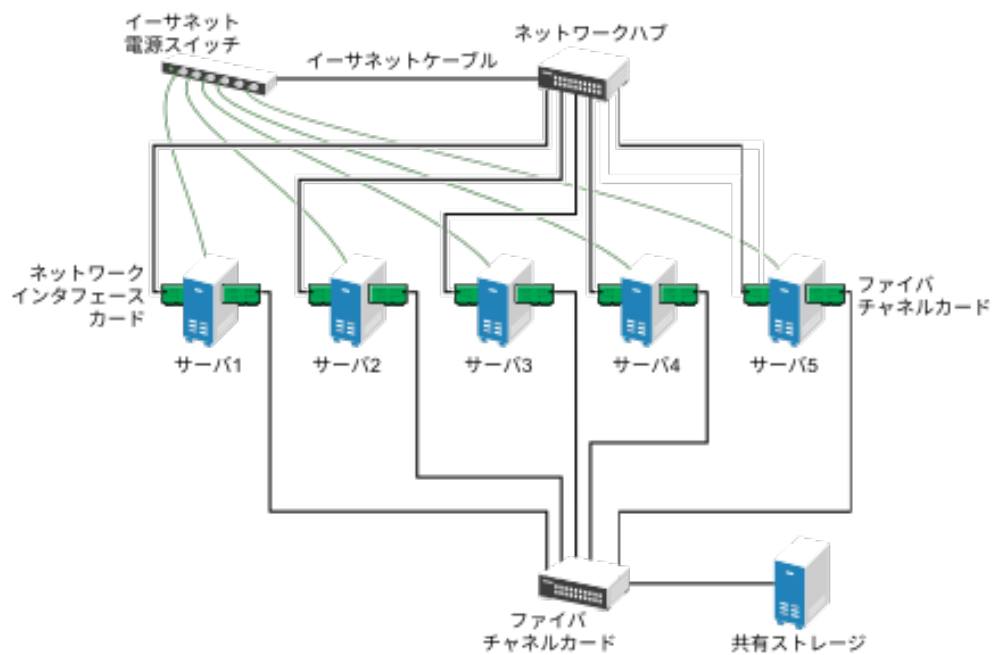


図 1.3: 一般的なファイバチャネルクラスタの設定

ファイバチャネルは最も高いパフォーマンスを提供しますが、iSCSIを利用するようにクラスタを設定することもできます。iSCSIは低コストなストレージエリアネットワーク(SAN)を作成するための方法として、ファイバチャネルの代わりに使用できます。次の図は、一般的なiSCSIクラスタの設定を表したものです。

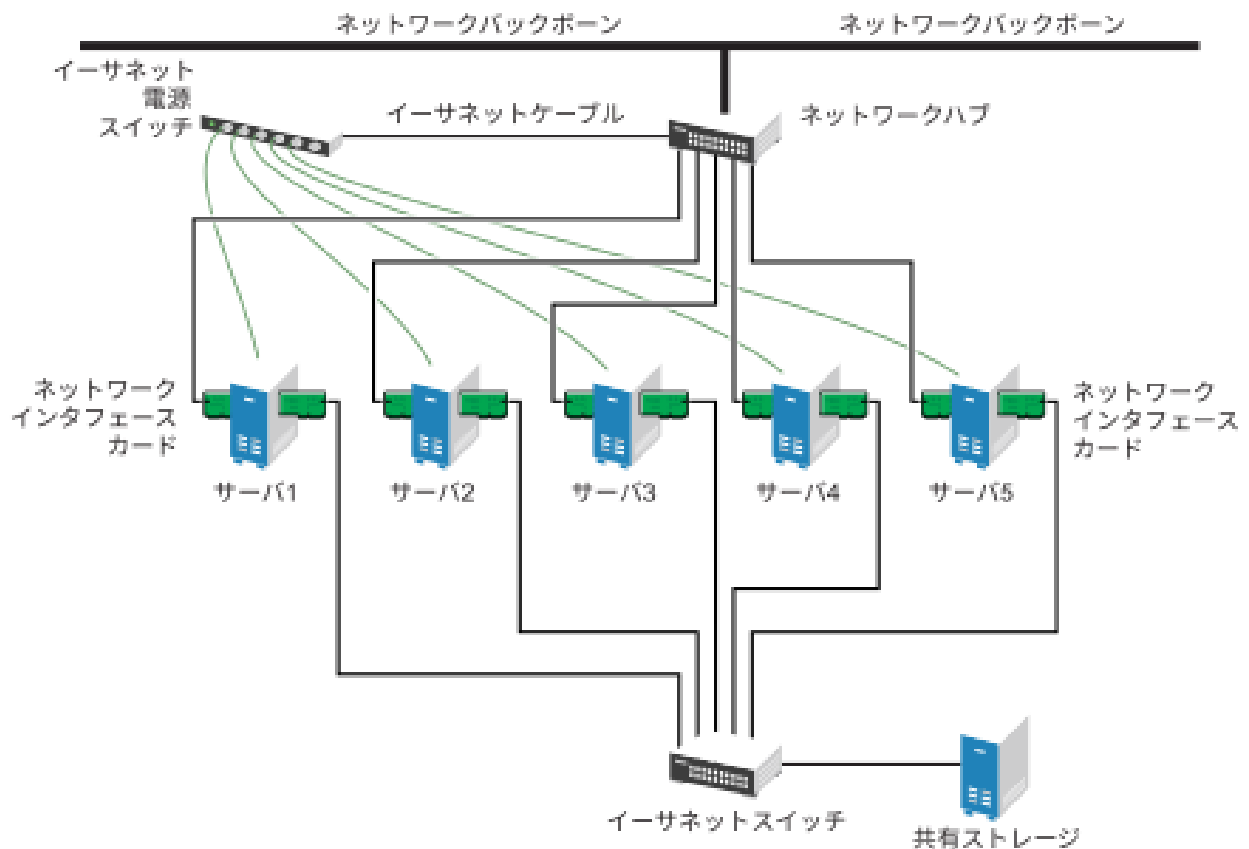


図 1.4: 一般的なiSCSIクラスタの設定

ほとんどのクラスタには共有ディスクサブシステムが含まれていますが、共有ディスクサブシステムなしのクラスタを作成することもできます。次の図は、共有ディスクサブシステムなしのクラスタを表したものです。

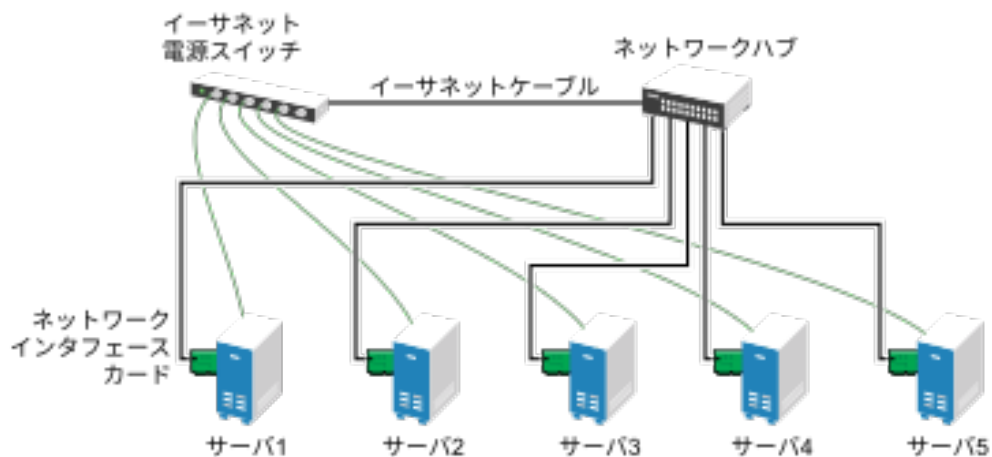


図 1.5: 共有ストレージなしの一般的なクラスタ設定

1.5 アーキテクチャ

このセクションでは、High Availability Extensionアーキテクチャの概要を説明します。アーキテクチャコンポーネントと、その相互運用方法について説明します。

1.5.1 アーキテクチャ層

High Availability Extensionのアーキテクチャは層化されています。図1.6「アーキテクチャ」に異なる層と関連するコンポーネントを示します。

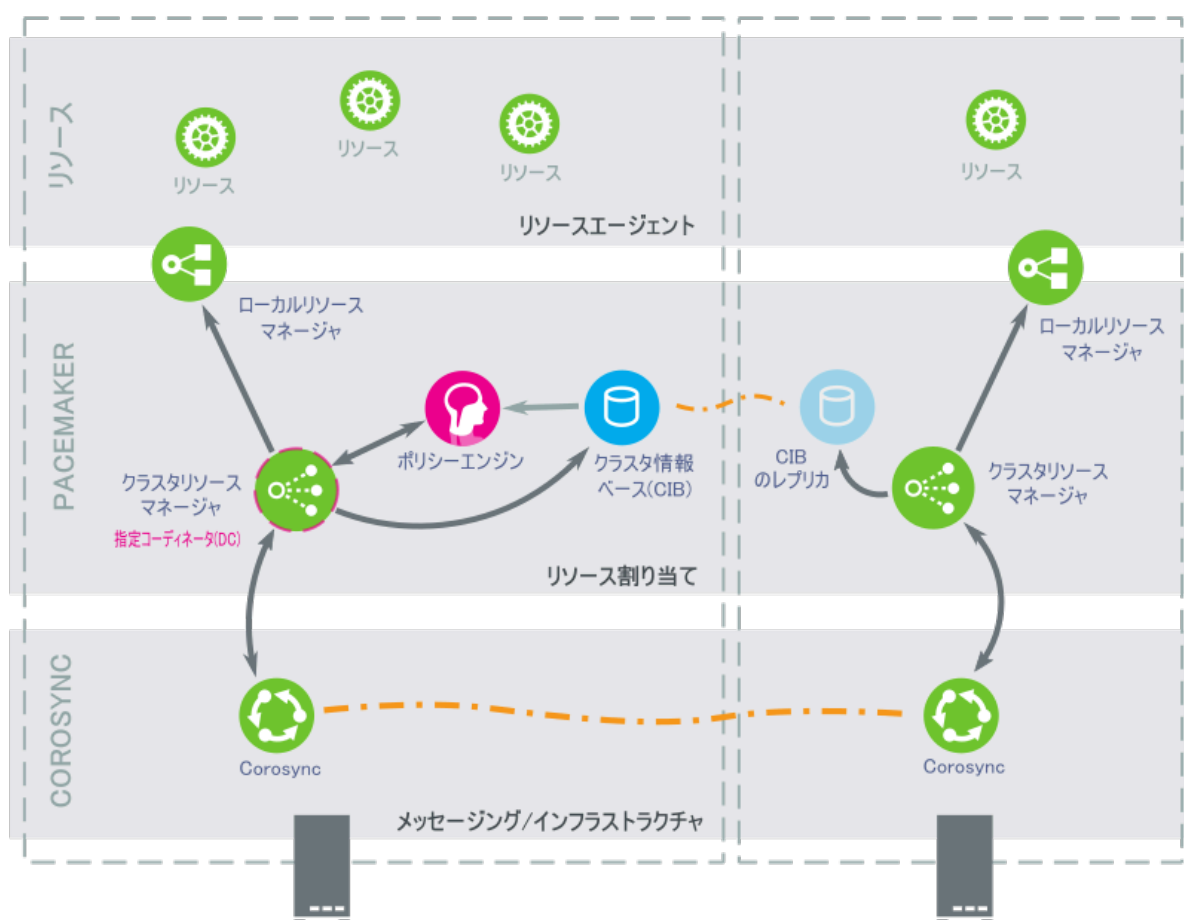


図 1.6: アーキテクチャ

1.5.1.1 メッセージングおよびインフラストラクチャ層

プライマリまたは最初の層は、メッセージングおよびインフラストラクチャの層で、Corosync層とも呼ばれます。この層には、「I am alive」信号やその他の情報を含むメッセージを送信するコンポーネントが含まれます。

1.5.1.2 リソース割り当て層

次の層はリソース割り当て層です。この層は最も複雑で、次のコンポーネントから設定されています。

CRM (クラスターリソースマネージャ)

リソース割り当て層のすべてのアクションは、クラスターリソースマネージャを通過します。リソース割り当て層の他のコンポーネント(または上位層のコンポーネント)による通信の必要性が発生した場合は、ローカルCRM経由で行います。すべてのノードで、CRMはCIB (クラスタ情報ベース)を維持しています。

CIB (クラスタ情報ベース)

クラスタ情報ベースは、メモリ内でクラスタ全体の設定や現在のステータスをXML形式で表すものです。すべてのクラスタオプション、ノード、リソース、制約、相互関係の定義が含まれます。CIBはすべてのクラスタノードへの更新の同期化も行います。指定コーディネータ(DC)が維持するマスターCIBがクラスタ内に1つあります。他のすべてのノードにはCIBのレプリカが含まれます。

指定コーディネータ(DC)

クラスタ内のCRMはDCとして選択されます。DCは、ノードのフェンシングやリソースの移動など、クラスタ全体におよぶ変更が必要かどうかを判断できる、クラスタ内で唯一のエンティティです。DCは、CIBのマスターコピーが保持されるノードでもあります。その他すべてのノードは、現在のDCから設定とリソース割り当て情報を取得します。DCは、メンバーシップの変更後、クラスタ内のすべてのノードから選抜されます。

PE (ポリシーエンジン)

指定コーディネータがクラスタ全体におよぶ変更を行う(新しいCIBに対応する)ことが必要になるたびに、ポリシーエンジンは現在の状態と設定に基づき、クラスタの次の状態を計算します。PEは(リソース)アクションのリストと、次のクラスタ状態に移るために必要な依存性を含む遷移グラフも作成します。PEは常にDC上で実行されます。

LRM(ローカルリソースマネージャ)

LRMはCRMに代わってローカルリソースエージェントを呼び出します(1.5.1.3項「リソース層」を参照)。そのため、操作の開始、停止、監視を行い、結果をCRMに報告します。LRMはそのローカルノード上のすべてのリソース関連情報の信頼できるソースです。

1.5.1.3 リソース層

最も上位の層はリソース層です。リソース層には1つ以上のリソースエージェント(RA)が含まれます。リソースエージェントは、一定の種類のサービス(リソース)を開始、停止、監視するために作成されたプログラム(通常はシェルスクリプト)です。リソースエージェントの呼び出しはLRMだけが行います。サードパーティはファイルシステム内の定義された場所に独自のエージェントを配置して、自社ソフトウェア用に、すぐに使えるクラスタ統合機能を提供することができます。

1.5.2 プロセスフロー

SUSE Linux Enterprise High Availability Extensionでは、PacemakerをCRMとして使用します。CRMは各クラスタノード上にインスタンスを持つデーモン(`crmd`)として実装されます。Pacemakerは、マスタとして動作する`crmd`インスタンスを1つ選択することにより、クラスタのすべての意思決定を一元化します。選択した`crmd`プロセス(またはその下のノード)で障害が発生したら、新しい`crmd`プロセスが確立されます。

クラスタの設定とクラスタ内のすべてのリソースの現在の状態を反映したCIBが、各ノードに保存されます。CIBのコンテンツはクラスタ全体で自動的に同期化されます。

クラスタ内で実行するアクションの多くは、クラスタ全体におよぶ変更を伴います。これらのアクションにはクラスタリソースの追加や削除、リソース制約の変更などがあります。このようなアクションを実行する場合は、クラスタ内でどのような変化が発生するのかを理解することが重要です。

たとえば、クラスタIPアドレスリソースを追加するとします。そのためには、コマンドラインツールかWebインタフェースを使用してCIBを変更できます。DC上でアクションを実行する必要はなく、クラスタ内の任意のノードでいずれかのツールを使用すれば、DCに反映されます。そして、DCがすべてのクラスタノードにCIBの変更を複製します。

CIBの情報に基づき、PEがクラスタの理想的な状態と実行方法を計算し、指示リストをDCに送ります。DCはメッセージング/インフラストラクチャ層を介してコマンドを送信し、他のノードの`crmd`ピアがこれらのコマンドを受信します。各`crmd`はLRM(`lrmd`として実装)を使用してリソースを変更します。`lrmd`はクラスタに対応しておらず、リソースエージェント(スクリプト)と直接通信します。

すべてのピアノードは操作結果をDCに返送します。DCが、すべての必要な操作がクラスタ内で成功したことを確認すると、クラスタはアイドル状態に戻り、次のイベントを待機します。予定通り実行されなかった操作があれば、CIBに記録された新しい情報を元に、PEを再度呼び出します。

場合によっては、共有データの保護や完全なリソース復旧のためにノードの電源を切らなければならないことがあります。このPacemakerにはフェンシングサブシステムとして`stonithd`が内蔵されています。STONITHは「Shoot The Other Node In The Head」の略です。通常は、STONITH共有ブロックデバイス、リモート管理ボード、またはリモートパワースイッチを使用して実装されま

す。Pacemakerで、STONITHデバイスはリソースとしてモデル化されており(そしてCIBで設定されており)、簡単に使用することができます。ただし、stonithdがSTONITHトポロジの把握を担うため、そのクライアントはノードのフェンシングを要求し、残りをstonithdが行います。

2 システム要件と推奨事項

次のセクションでは、SUSE® Linux Enterprise High Availability Extensionのシステム要件と前提条件について説明します。また、クラスタセットアップの推奨事項についても説明します。

2.1 ハードウェア要件

次のリストは、SUSE® Linux Enterprise High Availability Extensionに基づくクラスタのハードウェア要件を指定しています。これらの要件は、最低のハードウェア設定を表しています。クラスタの使用方法によっては、ハードウェアを追加しなければならないこともあります。

サーバ

2.2項「ソフトウェアの必要条件」に指定されたソフトウェアを搭載した1～32台のLinuxサーバ。サーバはベアメタルでも仮想マシンでも構いません。各サーバが同一のハードウェア設定(メモリ、ディスクスペースなど)になっている必要はありませんが、アーキテクチャは同じである必要があります。クロスプラットフォームのクラスタはサポートされていません。

`pacemaker_remote`を使用すると、32ノード制限を超えて追加のLinuxサーバを含めるようにクラスタを拡張できます。

通信チャンネル

クラスタノードあたり、少なくとも2つのTCP/IP通信メディア。ネットワーク機器は、クラスタ通信に使用する通信手段(マルチキャストまたはユニキャスト)をサポートする必要があります。通信メディアは100Mbit/s以上のデータレートをサポートする必要があります。サポートされるクラスタセットアップでは、2つ以上の冗長通信パスが必要です。これは次のように実行できます。

- ネットワークデバイスボンディング(推奨)。
- Corosync内の2つ目の通信チャンネル。
- インフラストラクチャ層のネットワークの耐障害性(ハイパーバイザーなど)。

詳細については、第13章「ネットワークデバイスボンディング」と手順4.3「冗長通信チャンネルの定義」をそれぞれ参照してください。

ノードフェンシング/STONITH

「スプリットブレイン」シナリオを回避するため、クラスタにはノードフェンシングメカニズムが必要です。スプリットブレインシナリオでは、クラスタノードは、お互いを認識していない2つ以上のグループに分割されます(ハードウェアまたはソフトウェアの障害か、ネットワーク接続の切断に

よる)。フェンシングメカニズムにより、問題のあるノードを分離します(通常はノードをリセットするか、ノードの電源をオフにすることによって分離します)。これをSTONITH (「Shoot the other node in the head」)と呼びます。ノードフェンシングメカニズムは、物理デバイス(電源スイッチ)でも、SBD (ディスクによるSTONITH)のようなメカニズムとウォッチドッグを組み合わせたものでも構いません。SBDを使用するには共有ストレージが必要です。

SBDが使用される場合を除き、High Availabilityクラスタの各ノードには少なくとも1つのSTONITHデバイスが必要です。ノードごとに複数のSTONITHデバイスを使用することを強くお勧めします。

❗ 重要: STONITHがない場合はサポートなし

- クラスタにはノードフェンシングメカニズムが必要です。
- グローバルクラスタオプション `stonith-enabled` および `startup-fencing` を `true` に設定する必要があります。これらを変更するとサポートされなくなります。

2.2 ソフトウェアの必要条件

クラスタに参加するすべてのノードに次のソフトウェアがインストールされている必要があります。

- SUSE® Linux Enterprise Server 12 SP5 (利用可能なすべてのオンラインアップデートが適用されていること)
- SUSE Linux Enterprise High Availability Extension 12 SP5 (利用可能なすべてのオンラインアップデートが適用されていること)
- (オプション)Geoクラスタの場合: Geo Clustering for SUSE Linux Enterprise High Availability Extension 12 SP5 (利用可能なすべてのオンラインアップデートが適用されていること)

2.3 ストレージ要件

一部のサービスでは、共有ストレージが必要です。外部NFS共有を使用する場合は、冗長通信パスを介してすべてのクラスタノードから確実にアクセスできる必要があります。

クラスタでデータの可用性を高めたい場合は、共有ディスクシステム(SAN: Storage Area Network)の利用をお勧めします。共有ディスクシステムを使用する場合は、次の要件を満たしていることを確認してください。

- メーカーの指示のに従い、共有ディスクシステムが適切に設定され、正しく動作していることを確認します。
- 共有ディスクシステム中のディスクを、ミラーリングまたはRAIDを使用して耐障害性が高められるように設定してください。
- 共有ディスクシステムのアクセスにiSCSIを使用している場合、iSCSIイニシエータとターゲットを正しく設定していることを確認します。
- 2台のマシンにデータを配分するミラーリングRAIDシステムを実装するためにDRBD*を使用する際、DRBDに提供されるデバイスにのみアクセスし、決してバックアップデバイスにはアクセスしないようにします。ボンディングNICを使用します。冗長性を確保するために、クラスタの残りの部分と同一のNICを利用できます。

SBDをSTONITHメカニズムとして使用する場合は、共有ストレージに対して追加の要件が適用されます。詳細については、[11.3項「要件」](#)を参照してください。

2.4 その他の要件と推奨事項

サポートされていて、役に立つHigh Availabilityセットアップについては、次の推奨事項を検討してください。

クラスタノード数

3つ以上のノードを持つクラスタに対して、奇数のクラスタノードを使用してクォーラムを持つようにすることを強くお勧めします。クォーラムの詳細については、[6.2項「クォーラムの判断」](#)を参照してください。

時刻同期

クラスタノードはクラスタ外のNTPサーバに同期する必要があります。詳細については、<https://documentation.suse.com/sles-12/html/SLES-all/cha-netz-xntp.html> を参照してください。

ノードが同期されていない場合、クラスタが正常に動作しないことがあります。また、同期が行われていないと、ログファイルとクラスタレポートの分析が非常に困難になります。ブートストラップスクリプトを使用するときにNTPがまだ設定されていない場合、警告が表示されます。

ネットワークインタフェースカード(NIC)名

すべてのノード上で同一である必要があります。

ホスト名およびIPアドレス

- 静的IPアドレスを使用します。
- `/etc/hosts` ファイルにあるすべてのクラスタノードを、完全修飾ホスト名およびショートホスト名で一覧表示します。クラスタのメンバーが名前で見つけられることが重要です。名前を使用できない場合、内部クラスタ通信は失敗します。
Pacemakerがノード名を取得する方法の詳細については、http://clusterlabs.org/doc/en-US/Pacemaker/1.1/html/Pacemaker_Explained/s-node-name.html も参照してください。

SSH

すべてのクラスタノードはSSHによって互いにアクセスできる必要があります。`crm report` (トラブルシューティング用)などのツールおよびHawk2の[履歴エクスプローラ]は、ノード間でパスワード不要のSSHアクセスを必要とします。それがない場合、現在のノードからしかデータを収集できません。



注記: 規定要件

パスワード不要のSSHアクセスが規定要件に適合しない場合は、[付録D rootアクセスなしでのクラスタレポートの実行](#)で説明されている次善策を使用して `crm report` を実行できます。

[履歴エクスプローラ]については、現在のところ、パスワード不要のログインに代わる方法はありません。

3 High Availability Extensionのインストール

初めてSUSE® Linux Enterprise High Availability Extensionを使用してHigh Availabilityクラスタを設定する場合、最も簡単な方法は、基本的な2ノードクラスタで開始することです。2ノードクラスタを使用して、一部のテストを実行することもできます。後で、AutoYaSTを使用して既存のクラスタノードのクローンを作成することにより、さらにノードを追加できます。クローンを作成したノードには、元のノードと同じパッケージがインストールされ、クローンノードは同じシステム設定を持つことになります。

前のバージョンのSUSE Linux Enterprise High Availability Extensionを実行する既存のクラスタをアップグレードする場合は、[第5章「クラスタアップグレードとソフトウェアパッケージの更新」](#)の章を参照してください。

3.1 手動インストール

High Availability Extension用のパッケージの手動インストールについては、項目「インストールおよびセットアップクイックスタート」を参照してください。

3.2 AutoYaSTによる大量インストールと展開

2ノードクラスタをインストールしてセットアップした後で、AutoYaSTを使用して既存のノードのクローンを作成し、クラスタにそのクローンを追加することによりクラスタを拡張できます。

AutoYaSTでは、インストールおよび設定データを含むプロファイルを使用します。このプロファイルによって、インストールする対象と、インストールしたシステムが最終的に使用準備が整ったシステムになるように設定する方法がAutoYaSTに指示されます。そこでこのプロファイルはさまざまな方法による大量配備に使用できます(たとえば、既存のクラスタノードのクローンなど)。



重要: 同一のハードウェアを使用している環境

[手順3.1「AutoYaSTによるクラスタノードのクローン作成」](#)では、同じハードウェア構成を持つ一群のマシンにSUSE Linux Enterprise High Availability Extension 12 SP5を展開していることを前提としています。

同じではないハードウェア上にクラスタノードを展開する必要がある場合は、『SUSE Linux Enterprise 12 SP5導入ガイド』、「Automated Installation」の章の「Rule-Based Autoinstallation」セクションを参照してください。

手順 3.1: AUTOYASTによるクラスタノードのクローン作成

1. クローンを作成するノードが正しくインストールされ、設定されていることを確認します。詳細については、SUSE Linux Enterprise High Availability Extension用の『インストールおよびセットアップクイックスタート』または第4章「YaSTクラスタモジュールの使用」を参照してください。
2. 単純な大量インストールについては、『SUSE Linux Enterprise 12 SP5 導入ガイド』の説明に従ってください。これには、次の基本ステップがあります。
 - a. AutoYaSTプロファイルの作成AutoYaST GUIを使用して、既存のシステム設定をもとにプロファイルを作成し、変更します。AutoYaSTでは、[高可用性]モジュールを選択し、[クローン]ボタンをクリックします。必要な場合は、他のモジュールの設定を調整し、その結果のコントロールファイルをXMLとして保存します。
DRBDを設定した場合、AutoYaST GUIでこのモジュールを選択してクローンを作成することもできます。
 - b. AutoYaSTプロファイルのソースと、他のノードのインストールルーチンに渡すパラメータを決定します。
 - c. SUSE Linux Enterprise ServerのソースとSUSE Linux Enterprise High Availability Extensionインストールデータを決定します。
 - d. 自動インストールのブートシナリオを決定し、設定します。
 - e. パラメータを手動で追加するか、または info ファイルを作成することにより、インストールルーチンにコマンド行を渡します。
 - f. 自動インストールプロセスを開始および監視します。

クローンのインストールに成功したら、次の手順を実行して、クローンノードをクラスタに加えます。

手順 3.2: クローンノードをオンラインにする

1. 4.5項「すべてのノードへの設定の転送」の説明に従って、Csync2を使用して、設定済みのノードからクローンノードへ重要な設定ファイルを転送します。
2. ノードをオンラインにするには、4.8項「クラスタをオンラインにする」の説明のとおり、クローンノード上でPacemakerサービスを開始します。

これで、`/etc/corosync/corosync.conf` ファイルがCsync2を介してクローンノードに適用されたので、クローンノードがクラスタに加わります。CIBは、クラスタノード間で自動的に同期されます。

4 YaSTクラスタモジュールの使用

YaSTクラスタモジュールでは、クラスタを手動で(最初から)設定するか、既存のクラスタのオプションを変更することができます。

ただし、クラスタの設定に自動化された方法を選ぶ場合は、項目「インストールおよびセットアップクイックスタート」を参照してください。このマニュアルでは、必要なパッケージのインストール方法と、`ha-cluster-bootstrap` スクリプトを使用して基本的な2ノードクラスタを設定する手順を説明しています。

たとえば、1つのノードをYaSTクラスタで設定してから、ブートストラップスクリプトの1つを使用して他のノードを統合させる(またはその逆も可能)など、両方のセットアップ方法を組み合わせることもできます。

4.1 用語の定義

YaSTクラスタモジュールおよびこの章で使用されているいくつかの主要な用語を以下に定義します。

バインドネットワークアドレス(`bindnetaddr`)

Corosyncエグゼクティブのバインド先のネットワークアドレス。クラスタ間の設定ファイルの共有を簡素化するため、Corosyncはネットワークインタフェースネットマスクを使用して、ネットワークのルーティングに使用されるアドレスビットのみをマスクします。たとえば、ローカルインタフェースが `192.168.5.92` でネットマスクが `255.255.255.0` の場合、`bindnetaddr` は `192.168.5.0` に設定します。ローカルインタフェースが `192.168.5.92` でネットマスクが `255.255.255.192` の場合は、`bindnetaddr` を `192.168.5.64` に設定します。



注記: すべてのノードのネットワークアドレス

すべてのノード上で同じCorosync設定が使用されるため、ネットワークアドレスは、特定のネットワークインタフェースのアドレスではなく、`bindnetaddr` として使用します。

`conntrack` ツール

カーネル内の接続トラッキングシステムとやり取りできるようにして、iptablesでのステートフルなパケット検査を可能にします。High Availability Extensionによって、クラスタノード間の接続ステータスを同期化するために使用されます。詳細については、<http://conntrack-tools.netfilter.org/> を参照してください。

Csync2

クラスタ内のすべてのノード、およびGeoクラスタ全体に設定ファイルを複製するために使用できる同期ツールです。Csync2は、同期グループ別にソートされた任意の数のホストを操作できます。各同期グループは、メンバーホストの独自のリストとその包含/除外パターン(同期グループ内でどのファイルを同期するか定義するパターン)を持っています。グループ、各グループに属するホスト名、および各グループの包含/除外ルールは、Csync2設定ファイル `/etc/csync2/csync2.cfg` で指定されます。

Csync2は、認証には、同期グループ内でIPアドレスと事前共有キーを使用します。管理者は、同期グループごとに1つのキーファイルを生成し、そのファイルをすべてのグループメンバにコピーする必要があります。

Csync2の詳細については、<http://oss.linbit.com/csync2/paper.pdf> を参照してください。

既存のクラスタ

「既存のクラスタ」という用語は、1つ以上のノードで構成されるクラスタを指すものとして使用されます。既存のクラスタは、通信チャネルを定義する基本的なCorosync設定を持ちますが、必ずしもリソース設定を持つとは限りません。

マルチキャスト

ネットワーク内で一対多数の通信に使用される技術で、クラスタ通信に使用できます。Corosyncはマルチキャストとユニキャストの両方をサポートしています。マルチキャストが会社のITポリシーに準拠しない場合、代わりにユニキャストを使用します。



注記: スイッチとマルチキャスト

クラスタ通信にマルチキャストを使用するには、ご使用のスイッチがマルチキャストをサポートしていることを確認します。

マルチキャストアドレス(`mcastaddr`)

Corosyncエグゼクティブによるマルチキャストに使用されるIPアドレス。このIPアドレスはIPv4またはIPv6のいずれかに設定できます。IPv6ネットワークを使用する場合は、ノードのIDを指定する必要があります。プライベートネットワークでは、どのようなマルチキャストアドレスでも使用できます。

マルチキャストポート(`mcastport`)

クラスタ通信に使用されるポート。Corosyncでは、マルチキャストの受信に指定する `mcastport` と、マルチキャストの送信に `mcastport -1` の、2つのポートを使用します。

冗長リングプロトコル(RRP)

ネットワーク障害の一部または全体に対する災害耐性のために、複数の冗長ローカルエリアネットワークが使用できるようになります。この方法では、ひとつのネットワークが作動中である限り、クラスタ通信を維持できます。Corosyncはトータム冗長リングプロトコルをサポートします。信頼できるソートされた方式でメッセージを配信するために、論理トークンパスリングがすべての参加ノードに課せられます。ノードがメッセージをブロードキャストできるのは、トークンを保持している場合のみです。

Corosyncに定義済みの冗長通信チャンネルを持つ場合、RRPを使用してこれらのインタフェースの使用方法をクラスタに伝えます。RRPでは次の3つのモードを使用できます(rrp_mode)。

- active に設定した場合、Corosyncは両方のインタフェースをアクティブに使用します。ただし、このモードは非推奨の機能です。
- passive に設定した場合、Corosyncは代わりに使用可能なネットワークを介してメッセージを送信します。
- none に設定した場合、RRPは無効になります。

ユニキャスト

ひとつのあて先ネットワークにメッセージを送信する技術Corosyncはマルチキャストとユニキャストの両方をサポートしています。Corosyncでは、ユニキャストはUDP-unicast (UDPU)として実装されます。

4.2 YaSTクラスタモジュール

YaSTを起動して、[高可用性] > [クラスタ]を選択します。または、コマンドラインでモジュールを開始します。

```
sudo yast2 cluster
```

次のリストは、YaSTクラスタモジュールで使用可能な画面の概要を示しています。この画面には、クラスタセットアップの成功に必要なパラメータが含まれているかどうか、またはそのパラメータがオプションであるかどうか説明されています。

通信チャンネル(必須)

クラスタノード間の通信に1つまたは2つの通信チャンネルを定義できます。転送プロトコルとして、マルチキャスト(UDP)またはユニキャスト(UDPU)のいずれかを使用します。詳細については、[4.3項「通信チャンネルの定義」](#)を参照してください。

！ 重要: 冗長通信パス

サポートされるクラスタセットアップでは、2つ以上の冗長通信パスが必要です。推奨される方法は、[第13章「ネットワークデバイスボンディング」](#)で説明されるように、ネットワークデバイスボンディングを使用することです。

使用できない場合は、Corosync内に2つ目の通信チャンネルを定義する必要があります。

セキュリティ(オプションだが推奨)

クラスタの認証設定を定義できます。共有シークレットが必要なHMAC/SHA1認証を使用して、メッセージを保護し、認証することができます。詳細については、[4.4項「認証設定の定義」](#)を参照してください。

Csync2の設定(オプションだが推奨)

Csync2では、設定変更を追跡して、クラスタノード間でファイルの同期を取ることができます。詳細については、[4.5項「すべてのノードへの設定の転送」](#)を参照してください。

conntrackdの設定(オプション)

ユーザスペース `conntrackd` を設定できます。iptablesでの「ステートフルな」パケット検査のためにconntrackツールを使用します。詳細については、[4.6項「クラスタノード間の接続ステータスの同期」](#)を参照してください。

サービス(必須)

クラスタノードをオンラインにするためにサービスを設定できます。ブート時にPacemakerサービスを開始するかどうか、およびノード間の通信に必要なポートをファイアウォールで開くかどうかを定義します。詳細については、[4.7項「サービスの設定」](#)を参照してください。

初めてクラスタモジュールを起動した場合は、モジュールが、ウィザードのように、基本設定に必要なすべてのステップをガイドします。そうでない場合は、左パネルのカテゴリをクリックして、ステップごとに設定オプションにアクセスします。



注記: YaSTクラスタモジュールの設定

YaSTクラスタモジュール内のいくつかの設定は、現在のノードにのみ適用されます。他の設定はCsync2を使用してすべてのノードに自動的に転送できます。これについての詳しい情報は次のセクションを参照してください。

4.3 通信チャネルの定義

クラスタノード間で正常な通信を行うには、少なくとも1つの通信チャネルを定義します。[手順 4.1](#)または[手順 4.2](#)のそれぞれで説明されるように、転送プロトコルとしてマルチキャスト(UDP)またはユニキャスト(UDPU)のいずれかを使用します。2番目の冗長チャネル([手順 4.3](#))を定義する場合は、両方の通信チャネルで「同じ」プロトコルを使用する必要があります。

YaST[通信チャネル]画面で定義されるすべての設定は、`/etc/corosync/corosync.conf`に書き込まれます。マルチキャストおよびユニキャストセットアップのサンプルファイルは、`/usr/share/doc/packages/corosync`にあります。

IPv4アドレスを使用する場合、ノードIDはオプションです。IPv6アドレスを使用する場合、ノードIDは必須です。各ノードにIDを手動で指定する代わりに、YaSTクラスタモジュールには、クラスタノードごとに固有のIDを自動的に生成するオプションが含まれています。

手順 4.1: 最初の通信チャネルの定義(マルチキャスト)

マルチキャストを使用する場合、すべてのクラスタノードに対して同じ `bindnetaddr`、`mcastaddr`、`mcastport` が使用されます。クラスタ内のすべてのノードは同じマルチキャストアドレスを使用することで互いを認識します。別のクラスタは、別のマルチキャストアドレスを使用します。

1. YaSTクラスタモジュールを起動して、[通信チャネル]カテゴリに切り替えます。
2. [転送]プロトコルを `Multicast` に設定します。
3. [バインドネットワークアドレス]を定義します。クラスタマルチキャストに使用するサブネットに値を設定します。
4. [マルチキャストアドレス]を定義します。
5. [ポート]を定義します。
6. クラスタノードごとに一意のIDを自動的に生成するには、[ノードIDの自動生成]を有効にしたままにします。
7. [クラスタ名]を定義します。
8. [期待する得票数]の数を入力します。これは、パーティションされたクラスタでCorosyncが**クォーラム**を計算する場合に重要です。デフォルトでは、各ノードには 1 票が割り当てられています。[期待する得票数]の数は、クラスタ内のノード数と一致する必要があります。
9. 変更内容を確認します。
10. 必要な場合は、[手順4.3「冗長通信チャネルの定義」](#)で説明するように、Corosyncで冗長な通信チャネルを定義します。

クラスタ - 通信チャンネル

トランスポート(T):
マルチキャスト

チャンネル
バインドネットワークアドレス(W): 192.168.1.0

冗長チャンネル(U) ☐
バインドネットワークアドレス(Q):

マルチキャストアドレス(S): 239.255.1.1

マルチキャストアドレス:

マルチキャストポート(M): 5405

マルチキャストポート(L):

メンバーアドレス:

IP	冗長IP	ノードID

追加(A) 削除(D) 編集(I)

クラスタ名(Q): NUE1

予想票数(X): 3

rrpモード(P): なし

☒ ノードIDの自動生成(G)

ヘルプ(H) 中止(R) 戻る(B) 次へ(N)

図 4.1: YASTクラスタ - マルチキャスト設定

クラスタ通信にマルチキャストではなくユニキャストを使用する場合は、次の手順に従います。

手順 4.2: 最初の通信チャンネルの定義(ユニキャスト)

1. YaSTクラスタモジュールを起動して、[通信チャンネル] カテゴリに切り替えます。
2. [転送] プロトコルを Unicast に設定します。
3. [ポート] を定義します。
4. ユニキャスト通信では、Corosyncはクラスタ内のすべてのノードのIPアドレスを認識する必要があります。クラスタの一部になる各ノードで、[追加] をクリックし、次の詳細を入力します。
 - [IPアドレス]
 - [冗長IPアドレス] (Corosyncで2つ目の通信チャンネルを使用する場合にのみ必要)
 - [ノードID] ([ノードIDの自動生成] オプションが無効になっている場合にのみ必要)

クラスタメンバーのアドレスを変更または削除するには、[編集]または[削除]ボタンを使用します。

5. クラスタノードごとに一意のIDを自動的に生成するには、[ノードIDの自動生成]を有効にしたままにします。
6. [クラスタ名]を定義します。
7. [期待する得票数]の数を入力します。これは、パーティションされたクラスタでCorosyncがクォーラムを計算する場合に重要です。デフォルトでは、各ノードには1票が割り当てられています。[期待する得票数]の数は、クラスタ内のノード数と一致する必要があります。
8. 変更内容を確認します。
9. 必要な場合は、[手順4.3「冗長通信チャネルの定義」](#)で説明するように、Corosyncで冗長な通信チャネルを定義します。

クラスタ - 通信チャネル

トランスポート(T):
ユニキャスト

チャンネル

バインドネットワークアドレス(W):
192.168.1.0

マルチキャストアドレス(S):
239.255.1.1

マルチキャストポート(M):
5405

☐ 冗長チャンネル(U)

バインドネットワークアドレス(O):

マルチキャストアドレス:

マルチキャストポート(L):

メンバーアドレス:

IP	冗長IP	ノードID
192.168.2.100		
192.168.2.101		
192.168.2.103		

追加(A)

削除(D)

編集(I)

クラスタ名(Q):
NUE1

予想票数(X):
3

rrpモード(P):
なし

☒ ノードIDの自動生成(Q)

ヘルプ(H)

中止(R)

戻る(B)

次へ(N)

図 4.2: YASTクラスタ - ユニキャスト設定

ネットワークデバイスボンディングが何らかの理由で使用できない場合、第2の選択は、Corosyncに冗長通信チャンネル(2つ目のリング)を定義することです。この方法では、2つの物理的に分かれたネットワークが通信に使用できます。1つのネットワークが失敗しても、クラスタノードは、もう一方のネットワークを介して通信できます。

Corosync内の追加の通信チャンネルは2つ目のトークンパスリングを形成します。`/etc/corosync/corosync.conf`では、設定した最初のチャンネルはプライマリリングで、ringnumber 0を取得します。2つ目のリング(冗長チャンネル)はringnumber 1を取得します。

Corosyncに定義済みの冗長通信チャンネルを持つ場合、RRPを使用してこれらのインタフェースの使用方法をクラスタに伝えます。RRPでは、2つの物理的に別個のネットワークが通信に使用されます。1つのネットワークが失敗しても、クラスタノードは、もう一方のネットワークを介して通信できます。

RRPでは次の3つのモードを使用できます。

- **active** に設定した場合、Corosyncは両方のインタフェースをアクティブに使用します。ただし、このモードは非推奨の機能です。
- **passive** に設定した場合、Corosyncは代わりに使用可能なネットワークを介してメッセージを送信します。
- **none** に設定した場合、RRPは無効になります。

手順 4.3: 冗長通信チャンネルの定義



重要: 冗長リングおよび/etc/hosts

Corosync内で複数のリングが設定されている場合、各ノードが複数のIPアドレスを持つことができます。これはすべてのノードの `/etc/hosts` に反映する必要があります。

1. YaSTクラスタモジュールを起動して、[通信チャンネル] カテゴリに切り替えます。
2. [冗長チャンネル] を有効にします。冗長チャンネルは、定義した最初の通信チャンネルと同じプロトコルを使用する必要があります。
3. マルチキャストを使用する場合は冗長チャンネル用に次のパラメータを入力します: 使用する[バインドネットワークアドレス]、[マルチキャストアドレス]、および[ポート]。
ユニキャストを使用する場合は次のパラメータを定義します: 使用する[バインドネットワークアドレス]、および[ポート]。クラスタに参加するすべてのノードのIPアドレスを入力します。
4. Corosyncに、異なるチャンネルを使用する方法とタイミングを伝えるには、使用する[rrp_mode]を選択します。

- 通信チャンネルが1つだけ定義されている場合、`[rrp-mode]`が自動的に無効化されます(値 なし)。
- `active`に設定した場合、Corosyncは両方のインタフェースをアクティブに使用します。ただし、このモードは非推奨の機能です。
- `passive`に設定した場合、Corosyncは代わりに使用可能なネットワークを介してメッセージを送信します。

RRPの使用時に、High Availability Extensionは現在のリングの状態を監視し、障害発生後に冗長リングを自動的に再度有効化します。

または、`corosync-cfgtool`を使用してリングの状態を手動で確認します。使用可能なオプションは `-h` で参照できます。

5. 変更内容を確認します。

4.4 認証設定の定義

クラスタの認証設定を定義するには、HMAC/SHA1認証を使用できます。共有シークレットが必要なHMAC/SHA認証を使用して、メッセージを保護し、認証する必要があります。指定した認証キー(パスワード)が、クラスタ中のすべてのノードで使用されます。

手順 4.4: 安全な認証を有効にする

1. YaSTクラスタモジュールを起動し、[セキュリティ]カテゴリに切り替えます。
2. [安全認証の有効化]をオンにします。
3. 新しく作成したクラスタの場合は、[認証キーファイルの生成]をクリックします。認証キーが作成され、`/etc/corosync/authkey`に書き込まれます。
ご使用のマシンを既存のクラスタに参加させたい場合、新しいキーファイルは生成しないでください。代わりに、いずれかのノードから `/etc/corosync/authkey` を(手動またはCsync2によって)ご使用のマシンにコピーします。
4. 変更内容を確認します。YaSTが設定を `/etc/corosync/corosync.conf` に書き込みます。



図 4.3: YASTクラスタ - セキュリティ

4.5 すべてのノードへの設定の転送

結果として生成された設定ファイルをすべてのノードに手動でコピーする代わりに、`csync2` ツールを使用して、クラスタ内のすべてのノードにレプリケートします。

これには、次の基本手順を必要とします。

1. YaSTによるCsync2の設定。
2. Csync2による設定ファイルの同期。

Csync2では、設定変更を追跡して、クラスタノード間でファイルの同期を取ることができます。

- 操作に対して重要なファイルのリストを定義できます。
- (他のクラスタノードに対して)これらのファイルの変更を表示できます。

- 1つのコマンドで複数の設定済みファイルの同期を取ることができます。
- `~/.bash_logout` の単純なシェルスクリプトを使用して、システムからログアウトする前に、同期化されていない変更について通知できます。

Csync2の詳細については、<http://oss.linbit.com/csync2/>と<http://oss.linbit.com/csync2/paper.pdf>にアクセスしてください。

4.5.1 YaSTによるCsync2の設定

1. YaSTクラスタモジュールを起動して、[Csync2] カテゴリに切り替えます。
2. 同期グループを指定するには、[同期ホスト] グループで [追加] をクリックし、クラスタ内のすべてのノードのローカルホスト名を入力します。ノードごとに、`hostname` コマンドから返された文字列を正確に使用する必要があります。



ヒント: ホスト名解決

ホスト名解決がネットワークで正しく機能しない場合は、各クラスタノードのホスト名とIPアドレスの組み合わせを指定することもできます。この指定には、`HOSTNAME@IP` 文字列 (たとえば、`alice@192.168.2.100`) を使用します。Csync2は、接続時にIPアドレスを使用します。

3. [事前共有キーの生成] をクリックして、同期グループのキーファイルを生成します。キーファイルは、`/etc/csync2/key_hagroup` に書き込まれます。このファイルは、作成後に、クラスタのすべてのメンバーに手動でコピーする必要があります。
4. すべてのノード間で、通常、同期される必要のあるファイルを [同期ファイル] リストに入れるには、[推奨ファイルの追加] をクリックします。
5. 同期するファイルのリストからファイルを [編集]、[追加]、または [削除] する場合は、該当する各ボタンを使用します。ファイルごとに絶対パス名を入力する必要があります。
6. [Csync2をオンにする] をクリックして、Csync2をアクティブにします。これによって次のコマンドが実行され、ブート時にCsync2が自動的に起動します。

```
root # systemctl enable csync2.socket
```

7. 変更内容を確認します。YaSTがCsync2の設定内容を `/etc/csync2/csync2.cfg` に書き込みます。

8. ここで同期プロセスを開始するには、4.5.2項「Csync2を使用した変更内容の同期」で続行します。



図 4.4: YASTクラスタ - CSYNC2

4.5.2 Csync2を使用した変更内容の同期

Csync2を使用してファイルを正常に同期するには、以下の前提条件が満たされている必要があります。

- 同じCsync2設定をすべてのクラスタノードで使用する必要があります。
- 同じCsync2認証キーをすべてのクラスタノードで使用する必要があります。
- Csync2はすべてのクラスタノード上で実行されている必要があります。

したがって、Csync2を初めて実行する前に、以下の準備を行う必要があります。

手順 4.5: CSYNC2による初期同期の準備

1. ファイル `/etc/csync2/csync2.cfg` を、4.5.1項「YaSTによるCsync2の設定」で説明されたとおりに設定した後、すべてのノードに手動でコピーします。
2. 4.5.1項のステップ 3の1つのノードで作成した `/etc/csync2/key_hagroup` ファイルを、クラスタ内のすべてのノードにコピーしてください。このファイルは、Csync2による認証で必要になります。ただし、すべてのノードで同じファイルでなければならないので、他のノードではファイルを再生成しないでください。
3. すべてのノード上で次のコマンドを実行して、Csync2サービスを今すぐ開始します。

```
root # systemctl start csync2.socket
```

手順 4.6: CSYNC2による設定ファイルの同期

1. 最初にすべてのファイルを一度同期させるには、設定の「コピー元」であるマシン上で次のコマンドを実行します。

```
root # csync2 -xv
```

これによって、すべてのファイルをその他のノードにプッシュすることで、一度に同期を行います。すべてのファイルが正常に同期されると、Csync2がエラーなしで終了します。同期対象の1つ以上のファイルが(現在のノードだけでなく)他のノード上で変更されている場合は、Csync2から衝突が報告されます。次の出力とよく似た出力が表示されます。

```
While syncing file /etc/corosync/corosync.conf:
ERROR from peer hex-14: File is also marked dirty here!
Finished with 1 errors.
```

2. 現在のノードのファイルバージョンが「最良」だと確信する場合は、そのファイルを強制して再同期を行い、競合を解決できます。

```
root # csync2 -f /etc/corosync/corosync.conf
root # csync2 -x
```

Csync2オプションの詳細については、次のコマンドを実行してください

```
csync2 -help
```



注記: 変更後の同期のプッシュ

Csync2は変更のみをプッシュします。Csync2はマシン間でファイルを絶えず同期しているわけではありません。

同期が必要なファイルを更新する際はいつも、変更を加えたマシン上で `csync2 -xv` を実行することで、変更をその他のマシンにプッシュする必要があります。変更されていないファイルが配置された他のマシン上でこのコマンドを実行しても、何も起こりません。

4.6 クラスタノード間の接続ステータスの同期

iptablesに対してステートフルなパケット検査ができるようにするには、conntrackツールを設定して使
用します。これには、次の基本手順を必要とします。

1. YaSTによるconntrackdの設定。
2. conntrackd (クラス: ocf、プロバイダ: heartbeat) のリソースの設定。Hawk2を使用する場合、Hawk2によって提案されるデフォルト値を使用します。

conntrackツールを設定したら、これをLinux Virtual Serverで使用できます。負荷バランスを参照し
てください。

手順 4.7: YASTによるconntrackdの設定

YaSTクラスタモジュールを使用して、ユーザスペース `conntrackd` を設定します。これには、そ
の他の通信チャンネルに使用されていない専用のネットワークインタフェースが必要です。デーモ
ンは後でリソースエージェントによって起動できます。

1. YaSTクラスタモジュールを起動して、[conntrackdの設定] カテゴリに切り替えます。
2. [専用インタフェース] を選択して、接続ステータスを同期します。選択したインタフェースのIPv4
アドレスが自動的に検出され、YaSTに表示されます。これはすでに設定済みで、マルチキャスト
をサポートしている必要があります。
3. 接続ステータスの同期に使用する[マルチキャストアドレス]を定義します。
4. [グループ番号]で、接続ステータスを同期させるグループのID番号を定義します。
5. [/etc/conntrackd/conntrackd.conf の生成]をクリックして、conntrackd の設定ファイルを
作成します。
6. 既存のクラスタでオプションを変更した場合、変更を確認して、クラスタモジュールを終了します。
7. クラスタ設定を先に進めるには、[次へ]をクリックして、4.7項「サービスの設定」で続行します。

クラスタ - conntrackdの設定

conntrackdは、クラスタノード間のファイアウォールステータスを複製できるデーモンです。
YaSTでは、conntrackdのいくつかの基本的な部分を設定できます。
conntrackdは、ocf:heartbeat:conntrackdで起動する必要があります。

専用インタフェース(D):

eth0

▼

IP: 10.161.11.176

マルチキャストアドレス(M):

239.180.93.156

グループ番号(G):

1

/etc/conntrackd/conntrackd.confを生成します

ヘルプ(H)

中止(R)

戻る(B)

次へ(N)

図 4.5: YASTクラスタ - conntrackd

4.7 サービスの設定

YaSTクラスタモジュールは、ブート時にノード上で一定のサービスを開始するかどうか定義します。サービスを手動で開始または停止するためにモジュールを使用することもできます。クラスタノードをオンラインにし、クラスタリソースマネージャを起動するには、Pacemakerをサービスとして実行する必要があります。

手順 4.8: PACEMAKERの有効化

1. YaSTクラスタモジュール内で、[サービス]カテゴリに切り替えます。
2. このクラスタノードがブートするたびにPacemakerを起動するには、[起動中]グループで該当するオプションを選択します。[起動中]グループで、[オフ]を選択する場合は、このノードがブートするたびに手動でPacemakerを起動する必要があります。Pacemakerを手動で起動するには、次のコマンドを使用します。

36

サービスの設定

SLE HA 12 SP5

```
root # systemctl start pacemaker
```

3. Pacemakerをただちに起動または停止するには、それぞれのボタンをクリックします。
4. 現在のマシン上でのクラスタ通信に必要なポートをファイアウォールで開くには、[ファイアウォールでポートを開く]をアクティブにします。この設定は、/etc/sysconfig/SuSEfirewall2.d/services/clusterに書き込まれます。
5. 変更内容を確認します。この設定は、すべてのクラスタノードではなく、ご使用のマシンにのみ適用されることにご注意ください。



図 4.6: YASTクラスタ - サービス

4.8 クラスタをオンラインにする

最初のクラスタ設定が完了した後、各クラスタノード上でPacemakerサービスを開始し、スタックをオンラインにします。

手順 4.9: **PACEMAKERの開始とその状態の確認**

1. 既存のノードにログインします。
2. サービスがすでに実行していることを確認します。

```
root # systemctl status pacemaker
```

実行されていない場合、Pacemakerをすぐに開始します。

```
root # systemctl start pacemaker
```

3. それぞれのクラスタノードに対してこの手順を繰り返します。
4. ノードの1つで、`crm status` コマンドを使用してクラスタの状態を確認します。すべてのノードがオンラインの場合、出力は次のようになります。

```
root # crm status
Last updated: Thu Jul  3 11:07:10 2014
Last change: Thu Jul  3 10:58:43 2014
Current DC: alice (175704363) - partition with quorum
2 Nodes configured
0 Resources configured

Online: [ alice bob ]
```

この出力は、クラスタリソースマネージャが起動し、リソースを管理できる状態にあることを示しています。

基本設定を完了し、ノードがオンラインになったら、クラスタリソースの設定を開始できます。crmシェル(crmsh)やHA Web Konsoleなどのクラスタ管理ツールのいずれかを使用します。詳細については、[第8章「クラスタリソースの設定と管理\(コマンドライン\)」](#)または[第7章「Hawk2を使用したクラスタリソースの設定と管理」](#)を参照してください。

5 クラスタアップグレードとソフトウェアパッケージの更新

この章では、次の2つの異なるシナリオ(SUSE Linux Enterprise High Availability Extensionの別バージョン(メジャーリリースまたはサービスパック)へのクラスタアップグレード、およびクラスタノード上の各パッケージの更新)について説明します。5.2項「最新の製品バージョンへのクラスタアップグレード」および5.3項「クラスタノード上のソフトウェアパッケージの更新」を参照してください。

クラスタをアップグレードする場合、アップグレードを開始する前に、5.2.1項「SLE HAおよびSLE HA Geoでサポートされるアップグレードパス」および5.2.2項「アップグレード前に必要な準備」を確認してください。

5.1 用語集

次に、この章で使用される最も重要な用語の定義を示します。

メジャーリリース、

一般出荷(GA)バージョン

SUSE Linux Enterprise (または任意のソフトウェア製品)のメジャーリリースとは、新しい機能やツールを導入する、非推奨になっていたコンポーネントを削除する、後方互換性のない変更が存在する、などの特徴を持った新バージョンです。

オフラインマイグレーション

新しい製品バージョンに後方互換性のない大幅な変更が含まれる場合、クラスタをオフラインマイグレーションでアップグレードする必要があります。つまり、すべてのノードをオフラインにしてクラスタ全体をアップグレードしてから、すべてのノードをオンラインに戻す必要があります。

ローリングアップグレード

ローリングアップグレードでは、一度に1つずつクラスタノードをアップグレードし、残りのクラスタは実行中のままにします。つまり、最初のノードをオフラインにしてアップグレードし、オンラインに戻してクラスタに参加させます。その後、すべてのクラスタノードがメジャーバージョンにアップグレードされるまで、1つずつアップグレードを続けます。

サービスパック(SP)

複数のパッチを組み合わせ、インストールまたは展開しやすい形式にします。サービスパックには番号が付けられ、通常、プログラムのセキュリティ修正、更新、アップグレード、または拡張機能が含まれます。

アップデート

パッケージの新しいマイナーバージョンのインストール。

アップグレード

パッケージまたは配布の新しい主要バージョンのインストール。これにより新機能がもたらされます。[オフラインマイグレーション](#)および[ローリングアップグレード](#)も参照してください。

5.2 最新の製品バージョンへのクラスタアップグレード

サポートされるアップグレードパスおよびアップグレードの実行方法は、現在の製品バージョンおよび移行したいターゲットバージョンの両方によって異なります。

- ローリングアップグレードは、製品バージョンのGAから次のサービスパックまで、および1つのサービスパックから次のサービスパックまでの間でのみサポートされます。
- あるメジャーバージョンから次のメジャーバージョン(たとえば、SLE HA11からSLE HA 12など)、または特定のメジャーバージョンに属するサービスパックから次のメジャーバージョン(たとえば、SLE HA 11SP3からSLE HA 12)へアップグレードするには、オフラインマイグレーションが必要です。

基本システム(SUSE Linux Enterprise Server)のアップグレードの詳細については、アップグレードするターゲットバージョンの『SUSE Linux Enterprise Server導入ガイド』を参照してください。このガイドは<https://documentation.suse.com/#sles/>で入手できます。

5.2.1項は、SLE HA (Geo)でサポートされているアップグレードパスの概要、および参照する追加のマニュアルを示しています。



重要: 混合クラスタおよびアップグレード後に元に戻す操作はサポートされない

- SUSE Linux Enterprise High Availability Extension 11/SUSE Linux Enterprise High Availability Extension 12で実行する混合クラスタはサポートされていません。
- 製品バージョン12へのアップグレードプロセス後に、製品バージョン11に戻す処理は、サポートされていません。


5.2.1 SLE HAおよびSLE HA Geoでサポートされるアップグレードパス

アップグレード元とアップグレード先	アップグレードパス	詳細情報の参照先
SLE HA 11 SP3から SLE HA (Geo) 12	オフラインマイグレーション	<ul style="list-style-type: none">● 基本システム: 『SUSE Linux Enterprise Server 12導入ガイド』の「Updating and Upgrading SUSE Linux Enterprise」のパート● SLE HA: クラスタ全体のオフラインマイグレーション● SLE HA Geo: 『Geo Clustering for SUSE Linux Enterprise High Availability Extension 12 Geo Clustering Quick Start』の「Upgrading from SLE HA (Geo) 11 SP3 to SLE HA Geo 12」の項
SLE HA (Geo) 11 SP4からSLE HA (Geo) 12 SP1	オフラインマイグレーション	<ul style="list-style-type: none">● 基本システム: 『SUSE Linux Enterprise Server 12 SP1導入ガイド』の「Updating and Upgrading SUSE Linux Enterprise」のパート● SLE HA: クラスタ全体のオフラインマイグレーション● SLE HA Geo: 『Geo Clustering for SUSE Linux Enterprise High Availability Extension 12 SP1 Geo Clustering Quick Start』の「Upgrading to the Latest Product Version」の項

アップグレード元とアップグレード先	アップグレードパス	詳細情報の参照先
SLE HA (Geo) 12からSLE HA (Geo) 12 SP1	ローリングアップグレード	<ul style="list-style-type: none"> ● 基本システム: 『SUSE Linux Enterprise Server 12 SP1導入ガイド』の「Updating and Upgrading SUSE Linux Enterprise」のパート ● SLE HA: クラスタ全体のローリングアップグレードの実行 ● SLE HA Geo: 『Geo Clustering for SUSE Linux Enterprise High Availability Extension 12 SP1 Geo Clustering Quick Start』の「Upgrading to the Latest Product Version」の項
SLE HA (Geo) 12 SP1からSLE HA (Geo) 12 SP2	ローリングアップグレード	<ul style="list-style-type: none"> ● 基本システム: 『SUSE Linux Enterprise Server 12 SP2導入ガイド』の「Updating and Upgrading SUSE Linux Enterprise」のパート ● SLE HA: クラスタ全体のローリングアップグレードの実行 ● SLE HA Geo: 『Geo Clustering for SUSE Linux Enterprise High Availability Extension 12 SP2 Geo Clustering Quick Start』の「Upgrading to the Latest Product Version」の項 ● DRBD 8から DRBD 9への移行DRBD 8から DRBD 9への移行

アップグレード元とアップグレード先	アップグレードパス	詳細情報の参照先
SLE HA (Geo) 12 SP2からSLE HA (Geo) 12 SP3	ローリングアップグレード	<ul style="list-style-type: none"> ● 基本システム: 『SUSE Linux Enterprise Server 12 SP3導入ガイド』の「Updating and Upgrading SUSE Linux Enterprise」のパート ● SLE HA: クラスタ全体のローリングアップグレードの実行 ● SLE HA Geo: 『Geo Clustering for SUSE Linux Enterprise High Availability Extension 12 SP3 Geo Clustering Guide』の「Upgrading to the Latest Product Version」の項
SLE HA (Geo) 12 SP3からSLE HA (Geo) 12 SP4	ローリングアップグレード	<ul style="list-style-type: none"> ● 基本システム: 『SUSE Linux Enterprise Server 12 SP4導入ガイド』の「Updating and Upgrading SUSE Linux Enterprise」のパート ● SLE HA: クラスタ全体のローリングアップグレードの実行 ● SLE HA Geo: 『Geo Clustering for SUSE Linux Enterprise High Availability Extension 12 SP4 Geo Clustering Guide』の「Upgrading to the Latest Product Version」の項

アップグレード元とアップグレード先	アップグレードパス	詳細情報の参照先
SLE HA (Geo) 12 SP4からSLE HA (Geo) 12 SP5	ローリングアップグレード	<ul style="list-style-type: none"> ● 基本システム: 『SUSE Linux Enterprise Server 12 SP5導入ガイド』の「Updating and Upgrading SUSE Linux Enterprise」のパート ● SLE HA: クラスタ全体のローリングアップグレードの実行 ● SLE HA Geo: 『Geo Clustering for SUSE Linux Enterprise High Availability Extension 12 SP5 Geo Clustering Guide』の「Upgrading to the Latest Product Version」の項

「詳細情報の参照先」の列に示すマニュアルはすべて<https://documentation.suse.com> で入手できます。

5.2.2 アップグレード前に必要な準備

バックアップ

システムバックアップが最新で、復元可能かどうかを確認します。

テスト

運用環境で実行する前に、まず、クラスタセットアップのステージングインスタンスでアップグレード手順をテストします。

これにより、メンテナンス期間に要するタイムフレームを予測できます。発生する可能性のある予期しない問題を検出し、解決するのに役立ちます。

5.2.3 オフラインマイグレーション

この項は次のシナリオに適用されます。

- SLE HA 11 SP3からSLE HA 12へのアップグレード
- SLE HA 12 SP4からSLE HA 11 SP1へのアップグレード

クラスタがまだこれらより古い製品バージョンに基づいている場合は、まず、必要なターゲットバージョンにアップグレードするためのソースとして使用できるバージョンのSUSE Linux Enterprise ServerおよびSUSE Linux Enterprise High Availability Extensionにクラスタをアップグレードします。High Availability Extension 12クラスタスタックでは、さまざまなコンポーネントに大幅な変更が導入されています(たとえば、`/etc/corosync/corosync.conf`、OCFS2のディスクフォーマットなど)。したがって、SUSE Linux Enterprise High Availability Extension 11バージョンからのローリングアップグレードはサポートされません。代わりに、[手順5.1「クラスタ全体のオフラインマイグレーション」](#)で説明されているように、すべてのクラスタノードをオフラインにしてから、クラスタ全体を移行する必要があります。

手順 5.1: クラスタ全体のオフラインマイグレーション

1. 各クラスタノードにログインし、次のコマンドを使用してクラスタスタックを停止します。

```
root # rcopenais stop
```

2. クラスタノードごとに、目的のターゲットバージョンのSUSE Linux Enterprise ServerおよびSUSE Linux Enterprise High Availability Extensionへのアップグレードを実行します。すでにGeoクラスタがセットアップされている場合、そのGeoクラスタをアップグレードするには、『Geo Clustering for SUSE Linux Enterprise High Availability Extension Geo Clustering Quick Start』に記載されている追加の指示を参照してください。個々のアップグレードプロセスの詳細を確認するには、[5.2.1項「SLE HAおよびSLE HA Geoでサポートされるアップグレードパス」](#)を参照してください。
3. アップグレードプロセスが完了した後で、SUSE Linux Enterprise ServerおよびSUSE Linux Enterprise High Availability Extensionのアップグレード済みバージョンがインストールされた各ノードを再起動します。
4. クラスタセットアップでOCFS2を使用する場合は、次のコマンドを実行して、オンデバイス構造をアップデートします。

```
root # o2cluster --update PATH_TO_DEVICE
```

SUSE Linux Enterprise High Availability Extension 12および12 SPxに付属しているアップデートされたOCFS12バージョンで必要とされるディスクに、パラメータが追加されます。

5. Corosyncバージョン2の `/etc/corosync/corosync.conf` をアップデートするには:

- a. 1つのノードにログインして、YaSTクラスタモジュールを起動します。

- b. [通信チャンネル] カテゴリに切り替えて、新しいパラメータ([クラスタ名]および[期待する得票数])の値を入力します。詳細については、[手順4.1「最初の通信チャンネルの定義\(マルチキャスト\)」](#)または[手順4.2「最初の通信チャンネルの定義\(ユニキャスト\)」](#)をそれぞれ参照してください。

Corosyncバージョン2で無効になっていた、見つからない他のオプションをYaSTが検出する場合、変更を求めるプロンプトが表示されます。

- c. YaSTで変更内容を確認します。YaSTが変更を `/etc/corosync/corosync.conf` に書き込みます。
- d. クラスタに対してCsync2が設定されている場合は、次のコマンドを使用して、アップデートされたCorosync設定を他のクラスタノードにプッシュします。

```
root # csync2 -xv
```

Csync2の詳細については、[4.5項「すべてのノードへの設定の転送」](#)を参照してください。または、`/etc/corosync/corosync.conf` をすべてのクラスタノードに手動でコピーして、更新されたCorosync設定の同期を取ります。

6. 各ノードにログインして、次のコマンドを使用してクラスタスタックを起動します。

```
root # systemctl start pacemaker
```

7. `crm status` またはHawk2を使用してクラスタの状態を確認します。

8. ブート時に起動するように以下のサービスを設定します。

```
root # systemctl enable pacemaker
root # systemctl enable hawk
root # systemctl enable sbd
```



注記: CIB構文バージョンのアップグレード

リソースをグループ化するタグと一部のACLの機能は、`pacemaker-2.0` 以上のCIB構文バージョンでのみ動作します(現在のバージョンを確認するには、`cibadmin -Q |grep validate-with` コマンドを使用します)。SUSE Linux Enterprise High Availability Extension 11 SPxからアップグレードした場合、デフォルトではCIBバージョンがアップグレード「されません」。最新のCIBバージョンに手動でアップグレードするには、以下のコマンドのいずれかを使用します:

```
root # cibadmin --upgrade --force
```

または

```
root # crm configure upgrade force
```

5.2.4 ローリングアップグレード

この項は次のシナリオに適用されます。

- SLE HA 12からSLE HA 12 SP1へのアップグレード
- SLE HA 12 SP1からSLE HA 12 SP2へのアップグレード
- SLE HA (Geo) 12 SP2からSLE HA (Geo) 12 SP3へのアップグレード
- SLE HA (Geo) 12 SP3からSLE HA (Geo) 12 SP4へのアップグレード
- SLE HA (Geo) 12 SP4からSLE HA (Geo) 12 SP5へのアップグレード



警告: アクティブなクラスタスタック

ノードのアップグレードを開始する前に、「そのノード上の」クラスタスタックを「停止」します。

ソフトウェアの更新中にノード上のクラスタリソースマネージャがアクティブな場合、アクティブノードのフェンシングのような予期しない結果を招く場合があります。

手順 5.2: クラスタ全体のローリングアップグレードの実行

1. アップグレードするノードで `root` としてログインし、クラスタスタックを停止します。

```
root # systemctl stop pacemaker
```

2. 目的のターゲットバージョンのSUSE Linux Enterprise ServerおよびSUSE Linux Enterprise High Availability Extensionへのアップグレードを実行します。個々のアップグレードプロセスの詳細を確認するには、[5.2.1項「SLE HAおよびSLE HA Geoでサポートされるアップグレードパス」](#)を参照してください。

3. アップグレードしたノードでクラスタスタックを再起動して、ノードをクラスタに再加入させます。

```
root # systemctl start pacemaker
```

4. 次のノードをオフラインにし、そのノードに関して手順を繰り返します。
5. `crm status` またはHawk2を使用してクラスタの状態を確認します。

！ 重要: ローリングアップグレードの時間制限

最新の製品バージョンに装備されている新機能は、「すべての」クラスタノードが最新の製品バージョンにアップグレードされた後でないと使用できません。混合バージョンのクラスタは、ローリングアップグレード中の短いタイムフレームでのみサポートされます。ローリングアップグレードは1週間以内に完了してください。

クラスタノードに異なるCRMバージョンが検出される場合、Hawk2の[状態]画面に警告も表示されません。

5.3 クラスタノード上のソフトウェアパッケージの更新

🚫 警告: アクティブなクラスタスタック

ノードの更新を開始する前に、クラスタスタックが影響を受けるか否かによって、「そのノード上の」クラスタスタックを「停止」するか、「ノードを保守モード」にします。詳細については、[ステップ 1](#)を参照してください。

ソフトウェアの更新中にノード上のクラスタリソースマネージャがアクティブな場合、アクティブノードのフェンシングのような予期しない結果を招く場合があります。

1. ノード上にパッケージ更新をインストールする前に、次の内容を確認してください。

- その更新は、SUSE Linux Enterprise High Availability ExtensionまたはGeo Clustering Extensionに属するパッケージに影響しますか。影響する場合は、ソフトウェアの更新を開始する前に、ノード上でクラスタスタックを停止します。

```
root # systemctl stop pacemaker
```

- パッケージ更新には再起動が必要ですか。必要な場合は、ソフトウェアの更新を開始する前に、ノード上でクラスタスタックを停止します。

```
root # systemctl stop pacemaker
```

- これらの状況のいずれにも該当しない場合は、クラスタスタックを停止する必要はありません。その場合は、ソフトウェアの更新を開始する前に、ノードを保守モードにします。

```
root # crm node maintenance NODE_NAME
```

保守モードの詳細については、16.2項「保守タスクのためのさまざまなオプション」を参照してください。

2. YaSTまたはZypperを使用してパッケージ更新をインストールします。

3. 更新が正常にインストールされたら、次のいずれかを行います。

- それぞれのノードでクラスタスタックを起動します(ステップ 1で停止した場合)。

```
root # systemctl start pacemaker
```

- または、ノードの保守フラグを削除して、ノードを通常モードに戻します。

```
root # crm node ready NODE_NAME
```

4. `crm status` またはHawk2を使用してクラスタの状態を確認します。

5.4 その他の情報

アップグレード先の製品の変更点と新機能の詳細については、それぞれのリリースノートを参照してください。リリースノートは、<https://www.suse.com/releasesnotes/>  で入手できます。

II 設定および管理

- 6 設定および管理の基本事項 51
- 7 Hawk2を使用したクラスタリソースの設定と管理 88
- 8 クラスタリソースの設定と管理(コマンドライン) 139
- 9 リソースエージェントの追加または変更 171
- 10 フェンシングとSTONITH 175
- 11 ストレージ保護とSBD 186
- 12 アクセス制御リスト 205
- 13 ネットワークデバイスボンディング 213
- 14 負荷バランス 219
- 15 Geoクラスタ(マルチサイトクラスタ) 232
- 16 保守タスクの実行 233

6 設定および管理の基本事項

HAクラスタの主な目的はユーザサービスの管理です。ユーザサービスの典型的な例は、Apache Webサーバまたはデータベースです。サービスとは、ユーザの観点からすると、指示に基づいて特別な何かを行うことを意味していますが、クラスタにとっては開始や停止できるリソースにすぎません。サービスの性質はクラスタには無関係なのです。

この章では、リソースを設定しクラスタを管理する場合に知っておく必要のある基本概念を紹介します。後続の章では、High Availability Extensionが提供する各管理ツールを使用して、主要な設定および管理タスクを行う方法を説明します。

6.1 ユースケースのシナリオ

一般的に、クラスタは次の2つのカテゴリのいずれかに分類されます。

- 2ノードクラスタ
- 2ノードより多いクラスタ。これは通常、奇数のノード数を意味します。

異なるトポロジを追加して、異なるユースケースを生成することもできます。次のユースケースは最も一般的です。

1つの場所の2ノードクラスタ

設定: FC SANまたは同様の共有ストレージ、レイヤ2ネットワーク。

使用シナリオ: サービスの高可用性、およびデータレプリケーションのデータ冗長性なしに焦点を当てた埋め込みクラスタ。このようなセットアップは無線ステーションや組立てラインコントローラなどに使用されます。

2つの場所の2ノードクラスタ(最も広く使用されている)

設定: 対称的なストレッチクラスタ、FC SAN、およびレイヤ2ネットワークのすべてが2つの場所に及ぶ。

使用シナリオ: サービスの高可用性、およびローカルデータの冗長性に焦点を当てた従来のストレッチクラスタ。データベースおよびエンタープライズリソース計画に適しており、ここ数年間で最も人気のあるセットアップの1つです。

3つの場所の奇数のノード数

設定: $2 \times N + 1$ ノード、FC SANが2つの主な場所に及ぶ。FC SANを使用しない補助的な3番目のサイト、過半数メーカーとして機能する。レイヤ2ネットワーク、少なくとも2つの主な場所に及ぶ。

使用シナリオ: サービスの高可用性、およびデータの冗長性に焦点を当てた従来のストレッチクラスタ。たとえば、データベース、エンタープライズリソースプランニング。

6.2 クォーラムの判断

1つ以上のノードとその他のクラスタ間で通信が失敗した場合は、常にクラスタパーティションが発生します。ノードは同じパーティション内の他のノードのみと通信可能で、切り離されたノードは認識しません。クラスタパーティションは、ノード(投票)の過半数を保有する場合、クォーラムを持つ(「定足数に達している」と定義されます。これを実現する方法は「クォーラム計算」によって実行されます。クォーラムはフェンシングの要件です。

クォーラム計算はSUSE Linux Enterprise High Availability Extension 11とSUSE Linux Enterprise High Availability Extension 12の間で変更されました。SUSE Linux Enterprise High Availability Extension 11では、クォーラムはPacemakerによって計算されました。SUSE Linux Enterprise High Availability Extension 12以降では、CorosyncがPacemakerの設定を変更せずに直接2ノードクラスタのクォーラムを処理できます。

クォーラムの計算方法は、次のような要因によって影響されます。

クラスタノード数

実行中のサービスを継続させるため、2ノードを超えるクラスタはクラスタパーティションの解決においてクォーラム(過半数)に依存します。次の数式に基づき、クラスタが機能するために必要な動作ノードの最少数を計算できます。

$$N \geq C/2 + 1$$

N = minimum number of operational nodes

C = number of cluster nodes

たとえば、5ノードクラスタでは、最低3つの動作ノード(または障害が発生する可能性のある2ノード)が必要です。

2ノードクラスタまたは奇数のクラスタノードのいずれかを使用することを強くお勧めします。2ノードクラスタは、2サイト間のストレッチセットアップで重要です。奇数のノード数を持つクラスタは、1つのシングルサイトで構築するか、または3つのサイト間で分散させることができます。

Corosyncの設定

Corosyncはメッセージングおよびメンバーシップ層です。[6.2.4項「2ノードクラスタのCorosync設定」](#)および[6.2.5項「NノードクラスタのCorosync設定」](#)を参照してください。

6.2.1 グローバルクラスタオプション

グローバルクラスタオプションは、一定の状況下でのクラスタの動作を制御します。それらは、セットにグループ化され、Hawk2や `crm` シェルなどのクラスタ管理ツールで表示したり、変更することができます。

事前に定義されている値は、通常は、そのまま保持できます。ただし、クラスタの主要機能を正しく機能させるには、クラスタの基本的なセットアップ後に、次のパラメータを調整する必要があります。

- グローバルオプション `no-quorum-policy`
- グローバルオプション `stonith-enabled`

6.2.2 グローバルオプション `no-quorum-policy`

このグローバルオプションは、クラスタパーティションにクォーラムがない(ノードの過半数がパーティションに含まれない)場合どうするかを定義します。

許容値は、次のとおりです。

ignore

`no-quorum-policy` を ignore に設定するとクラスタがクォーラムを持つように動作します。リソース管理は続行されます。

SLES 11では、この値が2ノードのクラスタ用の推奨設定でした。SLES 12以降、このオプションは廃止されました。設定と条件に基づいて、Corosyncはクラスタノードまたは単一ノードに「クォーラム」を与えます。または与えません。

2ノードのクラスタの場合、クォーラムが失われた場合の唯一の意味のある動作は、常に反応することです。最初のステップとして、クォーラムを失ったノードのフェンシングを試行してください。

freeze

クォーラムが失われた場合は、クラスタパーティションがフリーズします。リソース管理は続行されます。実行中のリソースは停止されません(ただし、イベントの監視に対応して再起動される可能性があります)。ただし、影響を受けたパーティション内では、以後のリソースが開始されません。一定のリソースが他のノードとの通信に依存しているクラスタの場合(たとえば、OCFS2マウントなど)は、この設定が推奨されます。この場合、デフォルト設定 `no-quorum-policy=stop` は、次のようなシナリオになるので有効ではありません。つまり、ピアノードが到達不能な間はそれらのリソースを停止できなくなります。その代わりに、停止の試行は最終的にタイムアウトし、stop failure になり、エスカレートされた復元とフェンシングを引き起こします。

stop (デフォルト値)

クォーラムが失われると、影響を受けるクラスタパーティション内のすべてのリソースが整然と停止します。

suicide

クォーラムが失われると、影響を受けるクラスタパーティション内のすべてのノードがフェンシングされます。このオプションは、SBDと組み合わせる場合にのみ機能します。第11章「ストレージ保護とSBD」を参照してください。

6.2.3 グローバルオプションstonith-enabled

このグローバルオプションは、フェンシングを適用して、STONITHデバイスによる、障害ノードや停止できないリソースを持つノードのダウンを許可するかどうか定義します。通常のクラスタ操作には、STONITHデバイスの使用が必要なので、このグローバルオプションは、デフォルトで true に設定されています。デフォルト値では、クラスタは、STONITHリソースが定義されていない場合にはリソースの開始を拒否します。

何らかの理由でフェンシングを無効にする必要がある場合は、stonith-enabled を false に設定しますが、これはご使用の製品のサポートステータスに影響を及ぼすことに注意してください。また、stonith-enabled="false" を指定すると、Distributed Lock Manager (DLM) のようなリソースやDLMによるすべてのサービス(cLVM、GFS2、OCFS2など)は開始できません。



重要: STONITHがない場合はサポートなし

STONITHがないクラスタはサポートされません。

6.2.4 2ノードクラスタのCorosync設定

ブートストラップスクリプトを使用する場合、Corosync設定には次のオプションを持つ quorum セクションがあります。

例 6.1: 2ノードクラスタのCOROSYNC設定の例

```
quorum {
    # Enable and configure quorum subsystem (default: off)
    # see also corosync.conf.5 and votequorum.5
    provider: corosync_votequorum
    expected_votes: 2
    two_node: 1
}
```

```
}
```

SUSE Linux Enterprise 11とは反対に、SUSE Linux Enterprise 12のvotequorumサブシステムは、Corosyncバージョン2.xで機能します。つまり、`no-quorum-policy=ignore` オプションは使用してはならないことを意味します。

デフォルトで、`two_node: 1` が設定されている場合、`wait_for_all` オプションが自動的に有効になります。`wait_for_all` が有効でない場合、クラスタは両方のノードで平行に開始される必要があります。または、最初のノードが、見つからない2番目のノードで起動フェンシングを実行します。

6.2.5 NノードクラスタのCorosync設定

2ノードクラスタを使用しない場合、Nノードクラスタに奇数のノードを使用することを強くお勧めします。クォーラム設定に関して、次のオプションがあります。

- `ha-cluster-join` コマンドを使用したノードの追加、または
- Corosync設定の手動調整。

`/etc/corosync/corosync.conf` を手動で調整する場合、次の設定を使用します。

例 6.2: NノードクラスタのCOROSYNC設定の例

```
quorum {  
    provider: corosync_votequorum ❶  
    expected_votes: N ❷  
    wait_for_all: 1 ❸  
}
```

- ❶ Corosyncからのクォーラムサービスの使用
- ❷ 予想される投票数。このパラメータは `quorum` セクション内で提供されるか、または `odelist` セクションが利用できる場合に自動的に計算されます。
- ❸ wait for all (WFA)機能を有効にします。WFAが有効な場合、クラスタはすべてのノードが認識可能になった後でのみ定足数に達します。一部の起動時の競合状態を回避するために、`wait_for_all` 設定を `1` に設定すると役立つ場合があります。たとえば、5ノードクラスタでは、すべてのノードに1つの投票が割り当てられているため、`expected_votes` を `5` に設定します。3つ以上のノードが互いに認識できる場合、クラスタパーティションが定足数に達し、動作を開始できます。

6.3 クラスタリソース

クラスタの管理者は、クラスタ内のサーバ上の各リソースや、サーバ上で実行する各アプリケーションに対してクラスタリソースを作成する必要があります。クラスタリソースには、Webサイト、電子メールサーバ、データベース、ファイルシステム、仮想マシン、およびユーザが常時使用できるようにする他のサーバベースのアプリケーションまたはサービスなどが含まれます。

6.3.1 リソース管理

リソースは、クラスタで使用する前にセットアップする必要があります。たとえば、Apacheサーバをクラスタリソースとして使用するには、まず、Apacheサーバをセットアップし、Apacheの環境設定を完了してから、クラスタで個々のリソースを起動します。

リソースに特定の環境要件がある場合は、それらの要件がすべてのクラスタノードに存在し、同一であることを確認してください。この種の設定は、High Availability Extensionでは管理されません。これは、管理者自身が行う必要があります。



注記: クラスタによって管理されるサービスには介入しないでください。

High Availability Extensionでリソースを管理しているときに、同じリソースを他の方法(クラスタ外で、たとえば、手動、ブート、再起動など)で開始したり、停止してはなりません。High Availability Extensionソフトウェアが、すべてのサービスの開始または停止アクションを実行します。

サービスがクラスタ制御下ですでに実行された後にテストまたは保守タスクを実行する必要がある場合は、リソース、ノード、またはクラスタ全体を保守モードに設定してから、これらのいずれかに手動でタッチしてください。詳細については、[16.2項「保守タスクのためのさまざまなオプション」](#)を参照してください。

クラスタ内でリソースを設定したら、クラスタ管理ツールを使用して、すべてのリソースを手動で起動、停止、クリーンアップ、削除、または移行します。これらの操作の詳細については、使用しているクラスタ管理ツールに応じて次のいずれかを参照してください。

- Hawk2: [第7章「Hawk2を使用したクラスタリソースの設定と管理」](#)
- crmsh: [第8章「クラスタリソースの設定と管理\(コマンドライン\)」](#)

6.3.2 サポートされるリソースエージェントクラス

追加するクラスタリソースごとに、リソースエージェントが準拠する基準を定義する必要があります。リソースエージェントは、提供するサービスを抽象化して正確なステータスをクラスタに渡すので、クラスタは管理するリソースについてコミットする必要がありません。クラスタは、リソースエージェントに依存して、start、stop、またはmonitorのコマンドの発行に適宜対応します。

通常、リソースエージェントはシェルスクリプトの形式で配布されます。High Availability Extensionは、次のクラスのリソースエージェントをサポートしています。

Open Cluster Framework (OCF)リソースエージェント

OCF RAエージェントは、High Availabilityでの使用に最適であり、特に、マルチステートリソースまたは特殊なモニタリング機能を必要とする場合に適しています。それらのエージェントは、通常、`/usr/lib/ocf/resource.d/provider`にあります。この機能はLSBスクリプトの機能と同様です。ただし、環境設定では、常に、パラメータの受け入れと処理を容易にする環境変数が使用されます。OCF仕様は<https://github.com/ClusterLabs/OCF-spec/blob/master/ra/1.0/resource-agent-api.md>で参照できます(リソースエージェントに関連するため)。OCF仕様には、アクション終了コードの厳密な定義があります。[9.3項「OCF戻りコードと障害回復」](#)を参照してください。クラスタは、それらの仕様に正確に準拠します。

すべてのOCFリソースエージェントは少なくとも `start`、`stop`、`status`、`monitor`、`meta-data` のアクションを持つ必要があります。`meta-data` アクションは、エージェントの設定方法についての情報を取得します。たとえば、プロバイダ `heartbeat` で `IPaddr` エージェントの詳細を知りたい場合は、次のコマンドを使用します。

```
OCF_ROOT=/usr/lib/ocf /usr/lib/ocf/resource.d/heartbeat/IPaddr meta-data
```

出力は、XML形式の情報であり、いくつかのセクションを含みます(一般説明、利用可能なパラメータ、エージェント用の利用可能なアクション)。

または、`crmsh`を使用して、OCFリソースエージェントに関する情報を表示します。詳細については、[8.1.3項「OCFリソースエージェントに関する情報の表示」](#)を参照してください。

Linux Standards Base (LSB)スクリプト

LSBリソースエージェントは一般にオペレーティングシステム/配布パッケージによって提供され、`/etc/init.d`にあります。リソースエージェントをクラスタで使用するには、それらのエージェントがLSB iniスクリプトの仕様に準拠している必要があります。たとえば、リソースエージェントには、いくつかのアクションが実装されている必要があります。それらのアクションとして、少なくとも `start`、`stop`、`restart`、`reload`、`force-reload`、`status` があります。詳細については、http://refspecs.linuxbase.org/LSB_4.1.0/LSB-Core-generic/LSB-Core-generic/iniscrptact.htmlを参照してください。

これらのサービスの構成は標準化されていません。High AvailabilityでLSBスクリプトを使用する場合は、該当のスクリプトの設定方法を理解する必要があります。これに関する情報は、多くの場合、`/usr/share/doc/packages/PACKAGENAME` 内の該当パッケージのマニュアルに記載されています。

Systemd

SUSE Linux Enterprise 12から、一般的なSystem V initデーモンがsystemdに置き代わりました。Pacemakerは、systemdサービスが存在する場合は、それを管理できます。initスクリプトの代わりに、systemdはユニットファイルを持ちます。一般的に、サービス(またはユニットファイル)は、オペレーティングシステムによって提供されます。既存のinitスクリプトを変換する場合は、<http://0pointer.de/blog/projects/systemd-for-admins-3.html> で詳細情報を検索してください。

サービス

現在、並列に存在する「通常」タイプのシステムサービスが多数あります: **LSB** (System V initに属する)、**systemd**、および(一部のディストリビューションでは) **upstart**。そのため、Pacemakerは、どれが指定のクラスターノードに適用されるのかをインテリジェントに理解する特殊なエイリアスをサポートします。これは、クラスターにsystemd、upstart、およびLSBサービスが混在する場合には特に役立ちます。Pacemakerは、次の順番で指定されたサービスを検索しようとしています: LSB (SYS-V) initスクリプト、systemdユニットファイル、またはUpstartジョブ。

Nagios

モニタリングプラグイン(かつてはNagiosプラグインと呼ばれていた)により、リモートホスト上のサービスを監視できます。Pacemakerは、モニタリングプラグインが存在する場合は、これを使用してリモートモニタリングを実行できます。詳細については、[6.6.1項「監視プラグインを使用したリモートホストでのサービスの監視」](#)を参照してください。

STONITH(フェンシング)リソースエージェント

このクラスは、フェンシング関係のリソース専用 사용됩니다。詳細については、[第10章「フェンシングとSTONITH」](#)を参照してください。

High Availability Extensionで提供されるエージェントは、OCF仕様に従って作成されています。

6.3.3 リソースのタイプ

次のリソースタイプを作成できます。

プリミティブ

プリミティブリソースは、リソースの中で最も基本的なタイプです。

選択したクラスタ管理ツールでプリミティブリソースを作成する方法については、次を参照してください。

- Hawk2: [手順7.5「プリミティブリソースの追加」](#)
- crmsh: [8.4.2項「クラスタリソースの作成」](#)

グループ

グループには、一緒の場所で見つけ、連続して開始し、逆の順序で停止する必要があるリソースセットが含まれます。詳細については、[6.3.5.1項「グループ」](#)を参照してください。

クローン

クローンは、複数のホスト上でアクティブにできるリソースです。対応するリソースエージェントがサポートしていれば、どのようなリソースもクローン化できます。詳細については、[6.3.5.2項「クローン」](#)を参照してください。

マルチステートリソース(旧称はマスタ/スレーブリソース)

マルチステートリソースは、クローンリソースの特殊なタイプで、複数のモードを持つことができます。詳細については、[6.3.5.3項「マルチステートリソース」](#)を参照してください。

6.3.4 リソーステンプレート

類似した設定のリソースを多く作成する最も簡単な方法は、リソーステンプレートを定義することです。定義された後でテンプレートは、プリミティブ内で参照したり、[6.5.3項「リソーステンプレートと制約」](#)で説明するように、特定のタイプの制約内で参照することができます。

プリミティブ内でテンプレートを参照すると、そのテンプレートで定義されている操作、インスタンス属性(パラメータ)、メタ属性、使用属性がすべてプリミティブに継承されます。さらに、プリミティブに対して特定の操作または属性を定義することもできます。これらのいずれかがテンプレートとプリミティブの両方で定義されていた場合、プリミティブで定義した値の方が、テンプレートで定義された値よりも優先されます。

選択したクラスタ管理ツールでリソーステンプレートを定義する方法については、次を参照してください。

- Hawk2: [手順7.6「リソーステンプレートの追加」](#)
- crmsh: [8.4.3項「リソーステンプレートの作成」](#)

6.3.5 高度なリソースタイプ

プリミティブは、最も単純なタイプのリソースなので、設定が容易ですが、クラスタ設定には、より高度なリソースタイプ(グループ、クローン、マルチステートリソースなど)が必要になることがあります。

6.3.5.1 グループ

クラスタリソースの中には、他のコンポーネントやリソースに依存しているものもあります。それぞれのコンポーネントやリソースが決められた順序で開始され、依存しているリソースと同じサーバ上で同時に実行していなければならない場合があります。この設定を簡素化するには、クラスタリソースグループを使用できます。

例 6.3: WEBサーバのリソースグループ

リソースグループの1例として、IPアドレスとファイルシステムを必要とするWebサーバがあります。この場合、各コンポーネントは、個々のリソースであり、それらが組み合わされてクラスタリソースグループを構成します。リソースグループは、1つ以上のサーバで実行されます。ソフトウェアまたはハードウェアが機能しない場合には、個々のクラスタリソースと同様に、グループはクラスタ内の別のサーバにフェールオーバーします。

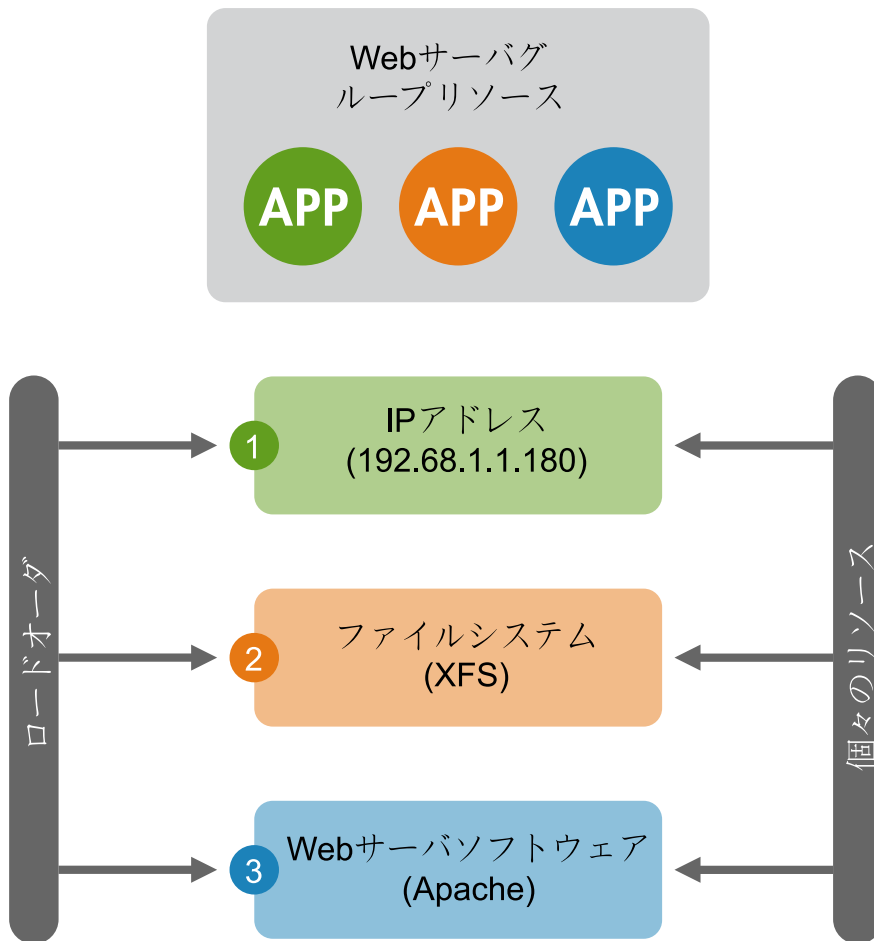


図 6.1: グループリソース

グループには次のプロパティがあります。

開始/停止

リソースは認識される順序で開始し、逆の順番で停止します。

依存関係

グループ内のリソースがどこかで開始できない場合は、グループ内のその後の全リソースは実行することができません。

コンテンツ

グループにはプリミティブクラスタリソースしか含むことができません。グループには1つ以上のリソースを含む必要があります。空の場合は設定は無効になります。グループリソースの子を参照するには、グループのIDではなく子のIDを使用します。

制約

制約でグループの子を参照することはできますが、通常はグループ名を使用することをお勧めします。

固着性

固着性はグループ内で統合可能なプロパティです。グループ内のアクティブな各メンバーは、グループの合計値に対して固着性を追加します。したがって、デフォルトの `resource-stickiness` が `100` で、グループに7つのメンバーがあり、そのうち5つがアクティブな場合は、グループが全体として、スコア `500` で、現在の場所を優先します。

リソース監視

グループのリソース監視を有効にするには、グループ内で監視の必要な各リソースに対して監視を設定する必要があります。

選択したクラスタ管理ツールでグループを作成する方法については、次を参照してください。

- Hawk2: [手順7.9「リソースグループを追加する」](#)
- crmsh: [8.4.10項「クラスタリソースグループの構成」](#)

6.3.5.2 クローン

クラスタ内の複数のノードで特定のリソースを同時に実行することができます。このためには、リソースをクローンとして設定する必要があります。クローンとして設定するリソースの一例として、OCFS2などのクラスタファイルシステムが挙げられます。提供されているどのリソースも、クローンとして設定できます。これは、リソースのリソースエージェントによってサポートされます。クローンリソースは、ホスティングされているノードによって異なる設定をすることもできます。

リソースクローンには次の3つのタイプがあります。

匿名クローン

最も簡単なクローンタイプです。実行場所にかかわらず、同じ動作をします。このため、マシンごとにアクティブな匿名クローンのインスタンスは1つだけ存在できます。

グローバルに固有なクローン

このリソースは独自のエントリです。1つのノードで実行しているクローンのインスタンスは、別なノードの別なインスタンスとは異なり、同じノードの2つのインスタンスが同一になることもありません。

ステートフルなクローン (マルチステートリソース)

このリソースのアクティブインスタンスは、アクティブとパッシブという2つの状態に分けられます。プライマリとセカンダリ、またはマスタとスレーブと呼ばれることもあります。ステートフルなクローンが、匿名またはグローバルに固有の場合もあります。[6.3.5.3項「マルチステートリソース」](#)も参照してください。

クローンは、グループまたは通常リソースを1つだけ含む必要があります。

リソースのモニタリングまたは制約を設定する場合、クローンには、単純なリソースとは異なる要件があります。詳細については、『Pacemaker Explained』(<http://www.clusterlabs.org/doc/>から入手可)を参照してください。特に、「Clones - Resources That Get Active on Multiple Hosts」のセクションを参照してください。

選択したクラスタ管理ツールでクローンを作成する方法については、次を参照してください。

- Hawk2: 手順7.10「クローンリソースの追加」
- crmsh: 8.4.11項「クローンリソースの設定」。

6.3.5.3 マルチステートリソース

マルチステートリソースは、クローンが得意とするところです。これにより、インスタンスを2つの動作モード(master または slave と呼ばれているが、任意の名前を割り当てることができる)のいずれかに設定できます。マルチステートリソースは、グループまたは通常リソースを1つだけ含む必要があります。

リソースの監視または制約を設定する場合、マルチステートリソースには、単純なリソースとは異なる要件があります。詳細については、『Pacemaker Explained』(<http://www.clusterlabs.org/doc/>から入手可)を参照してください。特に、「Multi-state - Resources That Have Multiple Modes」のセクションを参照してください。

6.3.6 リソースオプション(メタ属性)

追加した各リソースについて、オプションを定義できます。クラスタはオプションを使用して、リソースの動作方法を決定します。CRMに特定のリソースの処理方法を通知します。リソースオプションは、`crm_resource --meta` コマンドまたはHawk2を使用して設定できます(手順7.5「プリミティブリソースの追加」を参照)。

表 6.1: プリミティブリソースのオプション

オプション	説明	デフォルト
<u>優先度</u>	一部のリソースをアクティブにできない場合、クラスタは優先度の低いリソースを停止して、優先度の高いリソースをアクティブに維持します。	<u>0</u>

オプション	説明	デフォルト
<u>target-role</u>	クラスタが維持しようとするこのリソースの状態。使用できる値: <u>stopped</u> 、 <u>started</u> 、 <u>master</u>	<u>開始日</u>
<u>is-managed</u>	クラスタがリソースを開始して停止できるかどうか。使用できる値: <u>true</u> 、 <u>false</u> 値が <u>false</u> に設定される場合、リソースのステータスは依然として監視され、何らかの失敗が報告されます (<u>maintenance="true"</u> へのリソースの設定とは異なります)。	<u>true</u>
<u>保守モード</u>	リソースは手動でタッチできるかどうか。使用できる値: <u>true</u> 、 <u>false</u> <u>true</u> に設定すると、すべてのリソースが非管理対象になり、クラスタによる監視が停止されるため、ステータスは追跡されなくなります。クラスタによってクラスタリソースの再起動が試行される代わりに、ユーザがクラスタリソースを停止または再起動できます。	<u>false</u>
<u>resource-stickiness</u>	リソースが現在の状態をどの程度維持したいか。	計算済み
<u>migration-threshold</u>	ノードがこのリソースをホストできなくなるまで、このリソースについてノード上で発生する失敗の回数。	<u>INFINITY</u> (無効)
<u>multiple-active</u>	複数のノードでアクティブなリソースを検出した場合のクラスタの動作。使用でき	<u>stop_start</u>

オプション	説明	デフォルト
	る値: <u>block</u> (リソースを管理されていないとマークする)、 <u>stop_only</u> 、 <u>stop_start</u>	
<u>failure-timeout</u>	失敗が発生していないように動作する(リソースを失敗したノードに戻す)前に、待機する秒数	<u>0</u> (無効)
<u>allow-migrate</u>	<u>migrate_to</u> または <u>migrate_from</u> のアクションをサポートするリソースにリソース移行を許可。	<u>false</u>
<u>remote-node</u>	<p>このリソースが定義するリモートノードの名前。これにより、リモートノードのリソースが有効化されるだけでなく、リモートノードの識別に使用される固有の名前が定義されます。他のパラメータが設定されていない場合、この値はremote-portの接続先の<u>ホスト名とも</u> 見なされます。</p> <div>  <p>警告: 固有のIDの使用</p> <p>この値は、既存のリソースやノードIDとは重複させないでください。</p> </div>	なし(無効)
<u>remote-port</u>	pacemaker_remoteへのゲスト接続用のカスタムポート。	<u>3121</u>
<u>remote-addr</u>	リモートノードの名前がゲストのホスト名ではない場合に接続するIPアドレスまたはホスト名。	<u>remote-node</u> (ホスト名として使用される値)

オプション	説明	デフォルト
<code>remote-connect-timeout</code>	中断したゲスト接続がタイムアウトするまでの時間。	<code>60s</code>

6.3.7 インスタンス属性(パラメータ)

すべてのリソースクラスのスクリプトでは、動作方法および管理するサービスのインスタンスを指定するパラメータを指定できます。リソースエージェントがパラメータをサポートする場合、それらのパラメータを `crm_resource` コマンドまたは Hawk2 を使用して追加できます(手順7.5「プリミティブリソースの追加」を参照)。インスタンス属性は、`crm` コマンドラインユーティリティでは `params`、Hawk2 では `Parameter` と呼ばれます。OCF スクリプトでサポートされているインスタンス属性のリストは、次のコマンドを `root` として実行すると参照できます。

```
root # crm ra info [class:[provider:]]resource_agent
```

または(オプション部分なし):

```
root # crm ra info resource_agent
```

出力には、サポートされているすべての属性、それらの目的、およびデフォルト値が一覧されます。たとえば、次のコマンドを使用します。

```
root # crm ra info IPAddr
```

次の出力が返されます。

```
Manages virtual IPv4 addresses (portable version) (ocf:heartbeat:IPAddr)

This script manages IP alias IP addresses
It can add an IP alias, or remove one.

Parameters (* denotes required, [] the default):

ip* (string): IPv4 address
The IPv4 address to be configured in dotted quad notation, for example
"192.168.1.1".

nic (string, [eth0]): Network interface
The base network interface on which the IP address will be brought
online.

If left empty, the script will try and determine this from the
routing table.

Do NOT specify an alias interface in the form eth0:1 or anything here;
```

rather, specify the base interface only.

`cidr_netmask (string): Netmask`

The netmask for the interface in CIDR format. (ie, 24), or in dotted quad notation 255.255.255.0).

If unspecified, the script will also try to determine this from the routing table.

`broadcast (string): Broadcast address`

Broadcast address associated with the IP. If left empty, the script will determine this from the netmask.

`iflabel (string): Interface label`

You can specify an additional label for your IP address here.

`lvs_support (boolean, [false]): Enable support for LVS DR`

Enable support for LVS Direct Routing configurations. In case a IP address is stopped, only move it to the loopback device to allow the local node to continue to service requests, but no longer advertise it on the network.

`local_stop_script (string):`

Script called when the IP is released

`local_start_script (string):`

Script called when the IP is added

`ARP_INTERVAL_MS (integer, [500]): milliseconds between gratuitous ARPs`
milliseconds between ARPs

`ARP_REPEAT (integer, [10]): repeat count`

How many gratuitous ARPs to send out when bringing up a new address

`ARP_BACKGROUND (boolean, [yes]): run in background`

run in background (no longer any reason to do this)

`ARP_NETMASK (string, [ffffffffffff]): netmask for ARP`
netmask for ARP - in nonstandard hexadecimal format.

Operations' defaults (advisory minimum):

`start` `timeout=90`

`stop` `timeout=100`

`monitor_0` `interval=5s timeout=20s`



注記: グループ、クローン、またはマルチステートリソースのインスタンス属性

グループ、クローン、およびマルチステートリソースには、インスタンス属性がないので注意してください。ただし、インスタンス属性のセットは、グループ、クローン、またはマルチステートリソースの子によって継承されます。

6.3.8 リソース操作

デフォルトで、クラスタはリソースが良好な状態であることを保証しません。クラスタにこれを行わせるには、リソースの定義に監視操作を追加する必要があります。監視操作は、すべてのクラスまたはリソースエージェントに追加できます。詳細については、[6.4項「リソース監視」](#)を参照してください。

表 6.2: リソース操作のプロパティ

操作	説明
<u>id</u>	アクションに指定する名前。一意にする必要があります。(IDは表示されません)
<u>name</u>	実行するアクション。共通の値: <u>monitor</u> 、 <u>start</u> 、 <u>stop</u>
<u>interval</u>	操作を実行する頻度。単位: 秒
<u>timeout</u>	アクションが失敗したと宣言する前に待機する長さ。
<u>requires</u>	このアクションが発生する前に満たす必要のある条件。使用できる値: <u>nothing</u> 、 <u>quorum</u> 、 <u>fencing</u> デフォルトは、フェンシングが有効でリソースのクラスが <u>stonith</u> かどうかによります。STONITHリソースの場合、デフォルトは <u>nothing</u> です。

操作	説明
<u>on-fail</u>	<p>このアクションが失敗した場合に実行するアクション。使用できる値:</p> <ul style="list-style-type: none"> ● <u>ignore</u>: リソースが失敗しなかったのように動作します。 ● <u>block</u>: リソースにこれ以上の操作を実行しません。 ● <u>stop</u>: リソースを停止して、他の場所でも開始しません。 ● <u>restart</u>: リソースを停止して再起動します(別のノード上で)。 ● <u>fence</u>: リソースが失敗したノードを停止します(STONITH)。 ● <u>standby</u>: リソースが失敗したノードからすべてのリソースを移動させます。
<u>enabled</u>	<u>false</u> の場合、操作は存在していない場合と同様に処理されます。使用できる値: <u>true</u> 、 <u>false</u>
<u>role</u>	リソースにこの役割がある場合のみ操作を実行します。
<u>record-pending</u>	グローバルに設定したり、個々のリソースに対して設定できます。リソース上の「in-flight」操作の状態をCIBに反映させます。
<u>description</u>	操作について説明します。

6.3.9 タイムアウト値

リソースのタイムアウト値は次の3つのパラメータの影響を受けることがあります。

- op_defaults (操作のグローバルタイムアウト)
- リソーステンプレートに対して定義された特定のタイムアウト値
- リソースに対して定義された特定のタイムアウト値



注記: 値の優先度

リソースに対して「特定の」値が定義される場合、グローバルデフォルトより優先されます。また、リソースに対して定義された特定の値は、リソーステンプレートで定義された値より優先されます。

タイムアウト値を適切に設定することは非常に重要です。これらの値を短くしすぎると、次のような理由で、多数の(不必要な)フェンシング処理が発生します。

1. リソースでタイムアウトが発生すると、リソースは失敗し、クラスタはリソースを停止しようとします。
2. リソースの停止も失敗した場合(たとえば、停止のタイムアウト設定が短すぎる場合)、クラスタはノードをフェンシングします。クラスタは、このノードが制御できなくなっていると見なすからです。

操作に対するグローバルデフォルトを調整し、`crmsh`およびHawk2の両方で特定のタイムアウト値を設定できます。タイムアウト値の決定および設定のベストプラクティスは次のとおりです。

手順 6.1: タイムアウト値の決定

1. 負荷の下でリソースが開始および停止するためにかかる時間を確認します。
2. 必要に応じて op_defaults パラメータを追加し、それに応じて(デフォルト)タイムアウト値を設定します。
 - a. たとえば、op_defaults を 60 秒に設定します。

```
crm(live)configure# op_defaults timeout=60
```

- b. さらに長い時間を必要とするリソースについては、個別の値を定義します。
3. あるリソースに対して操作を設定する場合には、個別の start および stop 操作を追加します。Hawk2を使用して設定する場合、これらの操作に適したタイムアウト値候補が表示されます。

6.4 リソース監視

リソースが実行中であるかどうか確認するには、そのリソースにリソースの監視を設定しておく必要があります。

リソースモニタが障害を検出すると、次の処理が行われます。

- `/etc/corosync/corosync.conf` の `logging` セクションで指定された設定に従って、ログファイルメッセージが生成されます。
- 障害がクラスタ管理ツール(Hawk2、`crm status`)と、CIBステータスセクションに反映されます。
- クラスタが明瞭な復旧アクションを開始します。これらのアクションには、リソースを停止して障害状態を修復する、ローカルまたは別のノードでリソースを再起動するなどが含まれる場合があります。設定やクラスタの状態によっては、リソースが再起動されないこともあります。

リソースの監視を設定しなかった場合、開始成功後のリソース障害は通知されず、クラスタは常にリソース状態を良好として表示してしまいます。

停止されたリソースの監視

通常、リソースは動作している限り、クラスタのみによって監視されます。しかし、同時実行違反を検出するために、停止されるリソースの監視も設定する必要があります。次の例をご覧ください。

```
primitive dummy1 ocf:heartbeat:Dummy \  
    op monitor interval="300s" role="Stopped" timeout="10s" \  
    op monitor interval="30s" timeout="10s"
```

この設定は、300 秒ごとに、リソース `dummy1` に対する監視操作をトリガします。これは、リソースが `role="Stopped"` に入ると有効になります。実行中には、リソースは 30 秒ごとに監視されます。

プローブ

CRMはすべてのノードの各リソースに対して、probe と呼ばれる初期監視を実行します。probe はリソースのクリーンアップ後にも実行されます。1つのリソースに対して複数の監視操作が定義されている場合、CRMは最も時間間隔の短い監視を1つ選択し、そのタイムアウト値をプローブのデフォルトタイムアウトとして使用します。監視操作が何も設定されていない場合は、クラスタ規模のデフォルトが適用されます。デフォルトは、20 秒です(`op_defaults` パラメータの設定で別途指定されない場合)。自動計算や `op_defaults` の値に依存したくない場合は、このリソースの「プローブ」に対して特定の監視操作を定義します。 `interval` を 0 に設定した監視操作を追加することで、この操作を行います。たとえば次のようになります。

```
crm(live)configure# primitive rsc1 ocf:pacemaker:Dummy \  
    op probe interval=0 timeout=10s
```

```
op monitor interval="0" timeout="60"
```

`rsc1` のプローブは 60 秒でタイムアウトになります。この値は、`op_defaults` で定義されているグローバルタイムアウトや、その他の操作で設定されているタイムアウトとは無関係です。それぞれのリソースのプローブを指定するために `interval="0"` を設定していない場合、CRM は、そのリソースに定義されている監視操作がほかにはないかどうかを自動的に確認し、上で説明されているようにプローブのタイムアウト値を計算します。

選択したクラスタ管理ツールでリソースに対して監視操作を追加する方法については、次を参照してください。

- Hawk2: 手順7.13「操作の追加または変更」
- crmsh: 8.4.9項「リソース監視の設定」

6.5 リソースの制約

すべてのリソースを設定する以外にも、多くの作業が必要です。クラスタが必要なすべてのリソースを認識しても、正しく処理できるとは限りません。リソースの制約を指定して、リソースを実行可能なクラスタノード、リソースのロード順序、特定のリソースが依存している他のリソースを指定することができます。

6.5.1 制約のタイプ

使用可能な制約には3種類あります。

リソースの場所

場所の制約はリソースを実行できるノード、できないノード、または実行に適したノードを定義するものです。

リソースのコロケーション

コロケーション制約は、ノード上で一緒に実行可能な、または一緒に実行することが禁止されているリソースをクラスタに伝えます。

リソースの順序

アクションの順序を定義する、順序の制約です。

！ 重要: 制約および特定のタイプのリソースに関する制限

- リソースグループの「メンバー」に対してコロケーション制約を作成しないでください。代わりに、リソースグループ全体を指すリソース制約を作成してください。その他のタイプの制約はすべて、リソースグループのメンバーに対して使用しても問題ありません。
- クローンリソースまたはマルチステートリソースが適用されているリソースで制約を使用しないでください。制約はクローンまたはマルチステートリソースに適用する必要があり、その子リソースに適用することはできません。

6.5.1.1 リソースセット

6.5.1.1.1 制約を定義するためにリソースセットを使用する

場所、コロケーション、または順序の制約を定義するための別のフォーマットとして、`resource sets`を使用することができます。リソースセットでは、プリミティブが1つのセットでグループ化されます。以前は、これはリソースグループを定義するか(デザインを正確に表現できない場合もあった)、個々の制約として各関係を定義することでこの操作が可能でした。個々の制約として定義した場合、多数のリソースとの組み合わせが増えるにつれて、制約が飛躍的に増加しました。リソースセットを介した設定で、冗長性が常に低減されるわけではありませんが、次の例が示すように、定義内容の把握と管理がより容易になります。

例 6.4: 場所制約のリソースセット

たとえば、`crmsh`でリソースセット(`loc-alice`)の次の設定を使用して、2つの仮想IP (`vip1` および `vip2`)を同じノード、`alice`に配置できます。

```
crm(live)configure# primitive vip1 ocf:heartbeat:IPaddr2 params ip=192.168.1.5
crm(live)configure# primitive vip2 ocf:heartbeat:IPaddr2 params ip=192.168.1.6
crm(live)configure# location loc-alice { vip1 vip2 } inf: alice
```

リソースセットを使用してコロケーション制約の設定を置き換える場合は、次の2つの例を検討します。

例 6.5: コロケートされたリソースのチェーン

```
<constraints>
  <rsc_colocation id="coloc-1" rsc="B" with-rsc="A" score="INFINITY"/>
  <rsc_colocation id="coloc-2" rsc="C" with-rsc="B" score="INFINITY"/>
```

```
<rsc_colocation id="coloc-3" rsc="D" with-rsc="C" score="INFINITY"/>
</constraints>
```

リソースセットで表される同一の設定:

```
<constraints>
  <rsc_colocation id="coloc-1" score="INFINITY" >
    <resource_set id="colocated-set-example" sequential="true">
      <resource_ref id="A"/>
      <resource_ref id="B"/>
      <resource_ref id="C"/>
      <resource_ref id="D"/>
    </resource_set>
  </rsc_colocation>
</constraints>
```

リソースセットを使用して順序の制約の設定を置き換える場合は、次の2つの例を検討します。

例 6.6: 順序付けされたリソースのチェーン

```
<constraints>
  <rsc_order id="order-1" first="A" then="B" />
  <rsc_order id="order-2" first="B" then="C" />
  <rsc_order id="order-3" first="C" then="D" />
</constraints>
```

順序付けされたリソースを持つリソースセットを使用して、同様な目的を達成できます。

例 6.7: リソースセットとして表される順序付けされたリソースのチェーン

```
<constraints>
  <rsc_order id="order-1">
    <resource_set id="ordered-set-example" sequential="true">
      <resource_ref id="A"/>
      <resource_ref id="B"/>
      <resource_ref id="C"/>
      <resource_ref id="D"/>
    </resource_set>
  </rsc_order>
</constraints>
```

これらのセットは、順序付けされている (`sequential=true`) 場合もあれば、順序付けされていない場合 (`sequential=false`) 場合もあります。また、`require-all` 属性を使用して、`AND` および `OR` ロジック間を切り替えることができます。

6.5.1.1.2 依存関係のないコロケーション制約のリソースセット

同じノード上にリソースのグループを配置する方が役立つ場合があります(コロケーション制約を定義)、リソース間に困難な依存関係を持つことはありません。たとえば、同じノード上に2つのリソースを配置したいが、それらの一方で障害が発生した場合に他方をクラスタで再起動したくない場合があります。これは、`weak bond` コマンドを使用して、crmシェルで実行できます。

選択したクラスタ管理ツールでこれらの「弱い結合」を設定する方法については、次を参照してください。

- crmsh: 8.4.5.3項「依存性なしのリソースセットのコロケーション」

6.5.1.2 その他の情報

様々な種類の制約を追加する方法については、選択したクラスタ管理ツールに応じて次のいずれかを参照してください。

- Hawk2: 7.6項「制約の設定」
- crmsh: 8.4.5項「リソース制約の設定」

制約の設定の詳細や、順序付けおよびコロケーションの基本的な概念についての詳しいバックグラウンド情報は次のドキュメントを参照してください。これらのドキュメントは、<http://www.clusterlabs.org/doc/> で入手できます。

- 『Pacemaker Explained』の「Resource Constraints」の章
- 『Colocation Explained』
- 『オーダーの概要』

6.5.2 スコアと無限大

制約を定義する際は、スコアも扱う必要があります。あらゆる種類のスコアはクラスタの動作方法と密接に関連しています。スコアの操作によって、リソースのマイグレーションから、速度が低下したクラスタで停止するリソースの決定まで、あらゆる作業を実行できます。スコアはリソースごとに計算され、リソースに対して負のスコアが付けられているノードは、そのリソースを実行できません。リソースのスコアを計算した後、クラスタはスコアが最も高いノードを選択します。

`INFINITY` は現在 `1,000,000` と定義されています。この値の増減は、次の3つの基本ルールに従います。

- 任意の値+ INFINITY = INFINITY
- 任意の値- INFINITY = -INFINITY
- INFINITY - INFINITY = -INFINITY

リソース制約を定義する際は、各制約のスコアを指定します。スコアはこのリソース制約に割り当てる値を示します。スコアの高い制約は、それよりもスコアが低い制約より先に適用されます。1つのリソースに対して場所の制約を複数作成し、それぞれに異なるスコアを指定することで、リソースがフェールオーバーするノードの順序を指定できます。

6.5.3 リソーステンプレートと制約

リソーステンプレートを定義したら(6.3.4項「リソーステンプレート」を参照)、次のタイプの制約で参照できます。

- 順序の制約
- コロケーション制約
- rsc_ticket制約(Geoクラスタの場合)

ただし、コロケーション制約には、テンプレートへの参照を複数含めることはできません。リソースセットには、テンプレートへの参照を含めることはできません。

制約内で参照されたリソーステンプレートは、そのテンプレートから派生するすべてのプリミティブを表します。これは、そのリソーステンプレートを参照しているすべてのプリミティブリソースに、この制約が適用されることを意味します。制約内でリソーステンプレートを参照すれば、リソースセットの代替となり、クラスタ設定をかなりの程度単純化することができます。リソースセットの詳細については、[手順7.17「制約のためにリソースセットを使用する」](#)を参照してください。

6.5.4 フェールオーバーノード

リソースに障害が発生すると、自動的に再起動されます。現在のノードで再起動できない場合、または現在のノードでN回失敗した場合は、別のノードへのフェールオーバーが試行されます。リソースが失敗するたびに、その失敗回数が増加します。新しいノードへのマイグレートを行う基準(migration-threshold)となるリソースの失敗数を定義できます。クラスタ内に3つ以上ノードがある場合、特定のリソースのフェールオーバー先のノードはHigh Availabilityソフトウェアが選択します。

ただし、リソースに1つ以上の場所の制約とmigration-thresholdを設定することで、そのリソースのフェールオーバー先にするノードを指定できます。

選択したクラスタ管理ツールでフェールオーバーノードを指定する方法については、次を参照してください。

- Hawk2: 7.6.6項「リソースフェールオーバーノードの指定」
- crmsh: 8.4.6項「リソースフェールオーバーノードの指定」

例 6.8: マイグレーションしきい値 - プロセスフロー

たとえば、リソース「rsc1」に場所の制約を設定し、このリソースを「alice」で優先的に実行するように指定したと仮定します。そのノードで実行できなかった場合は、「migration-threshold」を確認して失敗回数と比較します。失敗回数 \geq マイグレーションしきい値の場合は、リソースは次の優先実行先として指定されているノードにマイグレートされます。

デフォルトでは、いったんしきい値に達すると、そのノードでは、リソースの失敗回数がリセットされるまで、失敗したリソースを実行できなくなります。これは、手動でクラスタ管理者が行うか、リソースに failure-timeout オプションを設定することで実行できます。

たとえば、migration-threshold=2 と failure-timeout=60s を設定すると、リソースは、2回の失敗の後に新しいノードに移行します。そして、1分後に復帰できます(固着性と制約のスコアによる)。

移行しきい値の概念には2つの例外があり、これらの例外は、リソースの開始失敗か、停止失敗のどちらかで発生します。

- 起動の失敗では、失敗回数が INFINITY に設定されるので、常に、即時に移行が行われます。
- 停止時の失敗ではフェンシングが発生します([stonith-enabled] がデフォルトである「true」に設定されている場合)。
STONITHリソースが定義されていない場合は(または stonith-enabled が false に設定されている場合)、リソースの移行は行われません。

選択したクラスタ管理ツールでマイグレーションしきい値を使用し、失敗回数をリセットする方法については、次を参照してください。

- Hawk2: 7.6.6項「リソースフェールオーバーノードの指定」
- crmsh: 8.4.6項「リソースフェールオーバーノードの指定」

6.5.5 フェールバックノード

ノードがオンライン状態に戻り、クラスタ内にある場合は、リソースが元のノードにフェールバックすることがあります。リソースを実行していたノードにリソースをフェールバックさせたくない場合や、リソースのフェールバック先として別のノードを指定する場合は、リソースの固着性の値を変更します。リソースの固着性は、リソースの作成時でも、その後でも指定できます。

リソース固着性値の指定時には、次の予想される結果について考慮してください。

0 の値:

デフォルトです。リソースはシステム内で最適な場所に配置されます。現在よりも「状態のよい」、または負荷の少ないノードが使用可能になると、移動することを意味しています。このオプションは自動フェールバックとほとんど同じですが、以前アクティブだったノード以外でもリソースをフェールバックできるという点が異なります。

0 より大きい値:

リソースは現在の場所に留まることを望んでいます、状態がよいノードが使用可能になると移動される可能性があります。値が大きくなるほど、リソースが現在の場所に留まることを強く望んでいることを示します。

0 より小さい値:

リソースは現在の場所から別な場所に移動することを望んでいます。絶対値が大きくなるほど、リソースが移動を強く望んでいることを示します。

INFINITY の値:

ノードがリソースの実行権利がなくなったために強制終了される場合(ノードのシャットダウン、ノードのスタンバイ、migration-thresholdに到達、または設定変更)以外は、リソースは常に現在の場所に留まります。このオプションは自動フェールバックを完全に無効にする場合とほとんど同じです。

-INFINITY の値:

リソースは現在の場所から常に移動されます。

6.5.6 負荷インパクトに基づくリソースの配置

すべてのリソースが同等ではありません。Xenゲストなどの一部のリソースでは、そのホストであるノードがリソースの容量要件を満たす必要があります。リソースの組み合わせられたニーズが提供された容量より大きくなるようにリソースが配置されると、リソースのパフォーマンスが低下します(あるいは失敗することさえあります)。

これを考慮に入れて、High Availability Extensionでは、次のパラメータを指定できます。

1. 一定のノードが提供する容量
2. 一定のリソースが要求する容量
3. リソースの配置に関する全体的なストラテジ

選択したクラスタ管理ツールでこれらの設定を設定する方法については、次を参照してください。

- Hawk2: 7.6.8項「[負荷インパクトに基づくリソース配置の設定](#)」
- crmsh: 8.4.8項「[負荷インパクトに基づくリソース配置の設定](#)」

ノードは、リソースの要件を満たすだけの空き容量があれば、そのリソースに対して資格があるとみなされます。High Availability Extensionにとって、容量の性質は重要ではありません。High Availability Extensionは、リソースをノードに移動する前に、リソースのすべての容量要件が満たされているかどうかを確認するだけです。

リソースの要件とノードが提供する容量を手動で設定するには、使用属性を使用します。使用属性に任意の名前を付け、設定に必要なだけ名前/値のペアを定義します。ただし、属性値は、整数にする必要があります。

使用属性を持つ複数のリソースがグループ化されていたり、これらにコロケーション制約がある場合、High Availability Extensionではそのことを考慮に入れます。可能な場合、これらのリソースは、すべての容量要件を満たすことができるノードに配置されます。



注記: グループの使用属性

リソースグループに対して使用属性を直接設定することはできません。ただし、グループの設定を簡素化するために、グループ内のすべてのリソースに必要な合計容量を含む使用属性を追加することができます。

High Availability Extensionには、ノードの容量とリソースの要件を自動的に検出し、設定する手段も用意されています。

NodeUtilization リソースエージェントは、ノードの容量をチェックします(CPUとRAMについて)。自動検出を設定するには、クラス、プロバイダ、タイプが `ocf:pacemaker:NodeUtilization` のクローンリソースを作成します。このクローンのインスタンスが各ノードに1つずつ実行している必要があります。インスタンスが開始すると、CIBでそのノードの設定にutilizationセクションが追加されます。

リソースの最小要件の自動検出(RAMとCPU)に配慮し、Xen リソースエージェントが改良されました。Xen リソースは、開始時点でRAMとCPUの消費状況を反映します。リソース設定には、使用属性が自動的に追加されます。



注記: Xenとlibvirtに異なるリソースエージェントを適用

ocf:heartbeat:Xen リソースエージェントは、libvirt に使用するべきではありません。libvirt ではマシン記述ファイルの変更が想定されているためです。

libvirt には、ocf:heartbeat:VirtualDomain リソースエージェントを使用します。

最小要件を検出することに加え、High Availability Extensionは、VirtualDomain リソースエージェントを通して現在の利用状況を監視することができ、仮想マシンでのCPUとRAMの使用状況を検出します。この機能を使用するには、クラス、プロバイダ、およびタイプがocf:heartbeat:VirtualDomain のリソースを設定します。次のインスタンス属性を使用できます: autoset_utilization_cpu および autoset_utilization_hv_memory。両方ともデフォルトは true です。これにより、監視サイクルのたびにCIBで使用値が更新されます。

容量と要件を手動と自動のどちらで設定する場合でも、placement-strategy プロパティ(グローバルクラスタオプション内)で、配置ストラテジを指定する必要があります。次の値を使用できます。

default (デフォルト値)

使用値は考慮しません。リソースは、場所のスコアに従って割り当てられます。スコアが同じであれば、リソースはノード間で均等に分散されます。

utilization

リソースの要件を満たすだけの空き容量がノードにあるかどうか決定する際に、利用率を確認します。ただし、負荷分散は、まだ、ノードに割り当てられたリソースの数に基づいて行われます。

minimal

リソースの要件を満たすだけの空き容量がノードにあるかどうか決定する際に、利用率を確認します。できるだけ少ない数のノードにリソースを集中しようとします(残りのノードの電力節約のため)。

balanced

リソースの要件を満たすだけの空き容量がノードにあるかどうか決定する際に、利用率を確認します。リソースを均等に分散して、リソースのパフォーマンスを最適化しようとします。



注記: リソース優先度の設定

使用できる配置ストラテジは、最善策であり、まだ、複雑なヒューリスティックソルバで、常に最適な割り当て結果を得るには至っていません。リソースの優先度を正しく設定して、最重要なリソースが最初にスケジュールされるようにしてください。

次の例は、同等のノードから成る3ノードクラスと4つの仮想マシンを示しています。

```
node alice utilization memory="4000"
node bob utilization memory="4000"
node charlie utilization memory="4000"
primitive xenA ocf:heartbeat:Xen utilization hv_memory="3500" \
    params xmfile="/etc/xen/shared-vm/vm1"
    meta priority="10"
primitive xenB ocf:heartbeat:Xen utilization hv_memory="2000" \
    params xmfile="/etc/xen/shared-vm/vm2"
    meta priority="1"
primitive xenC ocf:heartbeat:Xen utilization hv_memory="2000" \
    params xmfile="/etc/xen/shared-vm/vm3"
    meta priority="1"
primitive xenD ocf:heartbeat:Xen utilization hv_memory="1000" \
    params xmfile="/etc/xen/shared-vm/vm4"
    meta priority="5"
property placement-strategy="minimal"
```

3ノードはすべてアクティブであり、まず、リソース xenA がノードに配置され、次に、xenD が配置されます。xenB と xenC は、一緒に割り当てられるか、またはどちらか1つが xenD とともに割り当てられます。

1つのノードに障害が発生した場合、残りのノード上で利用できるメモリ合計が少なすぎて、これらのリソースすべてはホストできません。xenA は確実に割り当てられ、xenD も同様です。ただし、残りのリソース xenB と xenC は、そのどちらかしか割り当てられません。xenB と xenC の優先度は同等なので、結果はまだ決められません。これを解決するためにも、どちらかに高い優先度を設定する必要があります。

6.5.7 タグの使用によるリソースのグループ化

タグは最近Pacemakerに追加された新機能です。タグは、コロケーションの作成や関係の順序付けを行わずに、複数のリソースをただちに参照する方法です。これは、概念的に関連するリソースをグループ化するのに役立つ場合があります。たとえば、データベースに関連するいくつかのリソースがある場合、databases というタグを作成し、データベースに関連するすべてのリソースをこのタグに追加します。これにより、1つのコマンドでそれらすべてのリソースを停止または起動できます。

タグは制約でも使用できます。たとえば、次の場所制約 loc-db-prefer は、databases でタグ付けしたリソースのセットに適用されます。

```
location loc-db-prefer databases 100: alice
```

選択したクラスタ管理ツールでタグを作成する方法については、次を参照してください。

- Hawk2: 手順7.12「タグの追加」
- crmsh: 8.5.6項「リソースのグループ化/タグ付け」

6.6 リモートホストでのサービスの管理

リモートホストでサービスを監視および管理できることが、ここ数年の間にますます重要になってきています。SUSE Linux Enterprise High Availability Extension 11 SP3では、監視プラグインを紹介したリモートホスト上のサービスの詳細な監視機能を提供してきました。SUSE Linux Enterprise High Availability Extension 12 SP5では、最近追加された `pacemaker_remote` サービスを使用すると、リモートマシンにクラスタスタックをインストールしていなくても、実際のクラスタノードと同様にリモートホスト上のリソースを全面的に管理および監視できます。

6.6.1 監視プラグインを使用したリモートホストでのサービスの監視

仮想マシンの監視はVMエージェント(ハイパーバイザにゲストが出現する場合のみチェックを行う)を使用して行うか、VirtualDomainまたはXenエージェントから呼び出される外部スクリプトによって行うことができます。これまでは、精度の高い監視を行うには、仮想マシン内にHigh Availabilityスタックを完全にセットアップするしか方法がありませんでした。

今回、High Availability Extensionでは、監視プラグイン(旧称はNagiosプラグイン)に対するサポートを提供することで、リモートホスト上のサービスを監視できるようになりました。ゲストイメージを変更することなく、ゲストの外部ステータスを収集できます。たとえば、VMゲストはWebサービスまたは単純なネットワークリソースを実行している可能性があり、これらはアクセス可能である必要があります。Nagiosリソースエージェントによって、ゲスト上のWebサービスまたはネットワークリソースを監視できるようになりました。これらのサービスにアクセスできなくなった場合は、High Availability Extensionがそれぞれのゲストの再起動またはマイグレーションをトリガします。

ゲストがサービス(そのゲストによって使用されるNFSサーバなど)に依存している場合、そのサービスは、クラスタによって管理される通常のリソースか、Nagiosリソースによって監視される外部サービスのどちらかにすることができます。

Nagiosリソースを設定するには、ホスト上に次のパッケージをインストールする必要があります:

- monitoring-plugins
- monitoring-plugins-metadata

必要に応じて、YaSTまたはZypperが、これ以上のパッケージに対する依存性を解決します。

一般的な使用例としては、1つのリソースコンテナに属するリソースとして監視プラグインを設定します。このリソースコンテナは通常はVMです。いずれかのリソースに障害が発生したら、このコンテナが再起動されます。設定例については、[例6.10「監視プラグインのリソースの設定」](#)を参照してください。または、Nagiosリソースエージェントを使用してネットワーク経由でホストまたはサービスを監視する場合、このエージェントを通常のリソースとして設定することもできます。

例 6.10: 監視プラグインのリソースの設定

```
primitive vm1 ocf:heartbeat:VirtualDomain \
    params hypervisor="qemu:///system" config="/etc/libvirt/qemu/vm1.xml" \
    op start interval="0" timeout="90" \
    op stop interval="0" timeout="90" \
    op monitor interval="10" timeout="30"
primitive vm1-sshd nagios:check_tcp \
    params hostname="vm1" port="22" \ ❶
    op start interval="0" timeout="120" \ ❷
    op monitor interval="10"
group g-vm1-and-services vm1 vm1-sshd \
    meta container="vm1" ❸
```

- ❶ サポートされるパラメータは、監視プラグインの長いオプションと同じです。プラグインは、パラメータ `hostname` によってサービスと接続します。したがって、この属性の値は解決可能なホスト名かIPアドレスである必要があります。
- ❷ ゲストオペレーティングシステムが起動してサービスが実行されるまでには少し時間がかかるので、監視リソースの起動タイムアウトは十分な長さに設定する必要があります。
- ❸ タイプが `ocf:heartbeat:Xen`、`ocf:heartbeat:VirtualDomain`、または `ocf:heartbeat:lxc` のクラスタリソースコンテナ。VMまたはLinuxコンテナのいずれかに設定できます。

上の例には、`check_tcp` プラグイン用の1つのリソースしか含まれていませんが、様々なプラグインタイプ(たとえば、`check_http` や `check_udp` など)用に複数のリソースを設定することもできます。

複数のサービスのホスト名が同じである場合、`hostname` パラメータを個別のプリミティブに追加するのではなく、グループに対して指定することもできます。次に例を示します。

```
group g-vm1-and-services vm1 vm1-sshd vm1-httpd \
    meta container="vm1" \
    params hostname="vm1"
```

監視プラグインによって監視されているいずれかのサービスに、VM内で障害が発生した場合は、クラスタがこれを検出し、コンテナリソース(VM)を再起動します。この場合に実行される操作は、サービスの監視操作に関する `on-fail` 属性を指定することで設定できます。デフォルトでは、`restart-container` に設定されています。

VMのマイグレーションしきい値を検討する場合は、サービスの障害発生回数が考慮されます。

6.6.2 pacemaker_remoteを使用したリモートノードでのサービスの管理

pacemaker_remote サービスを使用すると、High Availability クラスタを仮想ノードまたはリモートベアメタルマシンに拡張することができます。クラスタスタックを実行して、クラスタのメンバーになる必要はありません。

High Availability Extensionでは現在、仮想環境(KVMおよびLXC)、およびこれらの仮想環境内に存在するリソースを起動できるようになりました(PacemakerまたはCorosyncの実行に仮想環境は必要としません)。

クラスタリソースとしての仮想マシンおよびVM内に存在するリソースの両方を管理する使用例では、次の設定を使用できるようになりました。

- 「通常」(ベアメタル)クラスタノードは、High Availability Extensionを実行します。
- 仮想マシンは、pacemaker_remote サービスを実行します(VM側で必要な設定はほとんどありません)。
- 「通常」クラスタノード上のクラスタスタックはVMを起動し、VM上で実行されている pacemaker_remote サービスに接続して、それらをリモートノードとしてクラスタに統合します。

リモートノードでクラスタスタックがインストールされていないときは、これには次の意味があります。

- リモートノードはクォーラムに参加しません。
- リモートノードはDCになることはできません。
- リモートノードは、スケーラビリティの制約に制限されません(Corosyncには32ノードのメンバー制限があります)。

remote_pacemaker サービスに関する詳細については(詳細な設定手順からなる複数の使用例を含む)、『Pacemaker Remote—Extending High Availability into Virtual Nodes』を参照してください(<http://www.clusterlabs.org/doc/> から入手可能)。

6.7 システムヘルスの監視

ノードがディスク容量が使い尽くしたために、そこに割り当てられたリソースを管理できなくなることを避けるため、High Availability Extensionでは、ocf:pacemaker:SysInfo というリソースエージェントが提供されています。これを使用して、ディスクパーティションに関してノードのヘルスを監視します。SysInfo RAは、#health_disk という名前のノード属性を作成します。この属性は、監視対象のディスク空き容量が指定された制限を下回ると red に設定されます。

ノードのヘルスがクリティカルな状態に達した場合のCRMの対応方法を定義するには、グローバルなクラスタオプションである `node-health-strategy` を使用します。

手順 6.2: システムヘルスの監視設定

ノードがディスク容量を使い尽くした場合に、リソースを自動的にノードから移動させるには、次の手順に従います。

1. `ocf:pacemaker:SysInfo` リソースを設定します。

```
primitive sysinfo ocf:pacemaker:SysInfo \  
  params disks="/tmp /var" ❶ min_disk_free="100M" ❷ disk_unit="M" ❸ \  
  op monitor interval="15s"
```

- ❶ 監視対象のディスクパーティション。たとえば、`/tmp`、`/usr`、`/var`、`/dev` など。複数のパーティションを属性値として指定するには、空白で区切ります。



注記: `/` のファイルシステムは常に監視されます。

`disks` でルートパーティション(`/`)を指定する必要はありません。これはデフォルトで常に監視されます。

- ❷ これらのパーティションの必要最小限の空きディスク容量。オプションで、計測に使用する単位を指定できます(上記の例では、メガバイトを表す `M` が使用されています)。指定しない場合、`min_disk_free` は `disk_unit` パラメータで定義されている単位にデフォルト設定されます。
 - ❸ ディスク容量をレポートする場合の単位。
2. リソース設定を完了するには、`ocf:pacemaker:SysInfo` のクローンを作成し、各クラスタノードでそれを起動します。
 3. `node-health-strategy` を `migrate-on-red` に設定します。

```
property node-health-strategy="migrate-on-red"
```

`#health_disk` 属性が `red` に設定されている場合、ポリシーエンジンによって、そのノードのリソースのスコアに `-INF` が追加されます。これにより、このノードからすべてのリソースが移動します。この処理はSTONITHリソースのところで停止しますが、STONITHリソースが実行されていない場合でも、ノードをフェンスすることができます。フェンスでCIBに直接アクセスすることで、動作を続行できるからです。

ノードのヘルス状態が **red** になったら、原因となる問題を解決します。次に **red** ステータスをクリアして、ノードを再びリソースの実行に適した状態にします。クラスターノードにログインして、次のいずれかの方法を使用します。

- 次のコマンドを実行します:

```
root # crm node status-attr NODE delete #health_disk
```

- 該当するノードでPacemakerを再起動します。
- ノードを再起動します。



ノードがサービスに復帰し、再びリソースを実行できるようになります。

6.8 その他の情報

<http://crmsh.github.io/> 

crmシェル(crmsh)、High Availabilityクラスター管理用の高度なコマンドラインインタフェースのホームページ。

<http://crmsh.github.io/documentation> 

crmshを使用した基本的なクラスター設定の『Getting Started』チュートリアルとcrmシェルの包括的なマニュアルを含む、crmシェルに関するいくつかのドキュメント。マニュアルは<http://crmsh.github.io/man-2.0/> で入手できます。チュートリアルは<http://crmsh.github.io/start-guide/> に用意されています。

<http://clusterlabs.org/> 

High Availability Extensionに含まれているクラスターリソースマネージャであるPacemakerのホームページ。

<http://www.clusterlabs.org/doc/> 

いくつかの包括的なマニュアルと一般的な概念を説明するより簡潔なドキュメント。次に例を示します。

- 『Pacemaker Explained』: 参考として包括的で詳細な情報が記載されています。
- 『Configuring Fencing with crmsh』: STONITHデバイスの設定方法および使用方法。
- 『Colocation Explained』
- 『オーダーの概要』

<https://clusterlabs.org> 

High Availability Linuxプロジェクトのホームページ。

7 Hawk2を使用したクラスタリソースの設定と管理

クラスタリソースを設定および管理する場合、HA Web Konsole (Hawk2)、またはcrmシェル(crmsh)コマンドラインユーティリティのいずれかを使用します。Hawkがインストールされている前のバージョンのSUSE® Linux Enterprise High Availability Extensionからアップグレードする場合は、パッケージが現在のバージョン、Hawk2で置き換えられます。

HawkのWebベースのユーザインタフェースを使用すれば、Linux以外のマシンから、Linuxクラスタを監視し、管理することができます。さらにこれは、ご使用のシステムが最小限のグラフィカルユーザインタフェースしか提供していない場合に最適なソリューションです。

7.1 Hawk2の要件

Hawk2にログインするには、次の要件を満たす必要があります。

- Hawk2で接続するすべてのクラスタノードに `hawk2` パッケージをインストールする必要があります。
- Hawk2を使用してクラスタノードにアクセスするマシンに必要なものは、JavaScriptとクッキーを有効にして接続を確立できるグラフィカルなWebブラウザです。
- Hawk2を使用するには、このWebインタフェースで接続するノード上で、それぞれのWebサービスが開始されている必要があります。詳細については、[手順7.1「Hawk2サービスの開始」](#)を参照してください。

`ha-cluster-bootstrap` パッケージからスクリプトを使用してクラスタをセットアップした場合、Hawk2サービスはすでに有効になっています。

- Hawk2ユーザは `haclient` グループのメンバーである必要があります。インストール時に `hacluster` という名前のLinuxユーザが作成されますが、このユーザが `haclient` グループに追加されます。セットアップ用に `ha-cluster-init` スクリプトを使用している場合は、`hacluster` ユーザに対してデフォルトパスワードが設定されます。

Hawk2を開始する前に、`hacluster` ユーザのパスワードを設定するか変更してください。または、`haclient` グループのメンバーである新しいユーザを作成してください。

Hawk2を使用して接続する各ノードでこれを実行します。

手順 7.1: HAWK2サービスの開始

1. 接続先にするノードで、シェルを開き、rootとしてログインします。
2. 次のように入力して、サービスのステータスをチェックします。

```
root # systemctl status hawk
```

3. サービスが実行されていない場合は、次のコマンドでサービスを開始します。

```
root # systemctl start hawk
```

ブート時にHawk2を自動的に起動したい場合は、次のコマンドを実行します。

```
root # systemctl enable hawk
```

7.2 ログイン

Hawk2 Webインタフェースは、HTTPSプロトコルとポート 7630 を使用します。

Hawk2を使用して個々のクラスタノードにログインする代わりに、浮動、仮想IPアドレス(IPaddr または IPaddr2)をクラスタリソースとして設定できます。そのための特別な設定は不要です。サービスがどの物理ノードで実行されていても、クライアントはHawkサービスに接続できます。

クラスタを ha-cluster-bootstrap スクリプトを使用して設定する際には、クラスタ管理用に仮想IPを設定するかどうかを求められます。

手順 7.2: HAWK2 WEBインタフェースへのログイン

1. いずれかのマシンでWebブラウザを起動し、次のURLを入力します。

```
https://HAWKSERVER:7630/
```

HAWKSERVER の部分は、Hawk Webサービスを実行するクラスタノードのIPアドレスまたはホスト名で置き換えます。Hawk2でクラスタ管理用に仮想IPアドレスを設定した場合、その仮想IPアドレスで HAWKSERVER を置き換えます。



注記: 証明書の警告

初めてURLにアクセスしようとするときに証明書の警告が表示される場合は、自己署名証明書が使用されています。自己署名証明書は、デフォルトでは信頼されません。

証明書を検証するには、クラスタオペレータに証明書の詳細を求めます。

続行するには、ブラウザに例外を追加して警告をバイパスします。

自己署名証明書を公式認証局によって署名された証明書で置き換える方法の詳細については、[自己署名証明書の置き換え](#)を参照してください。

2. Hawk2ログイン画面で、`hacluster` ユーザ(または、`haclient` グループのメンバーである他の任意のユーザ)の[ユーザ名]と[パスワード]を入力します。
3. [ログイン]をクリックします。

7.3 Hawk2の概要: 主な構成要素

Hawk2にログインすると、左側にはナビゲーションバー、右側には複数のリンクが含まれる最上位の行が表示されます。



注記: Hawk2で利用できる機能

デフォルトでは、`root` または `hacluster` としてログインしたユーザは、すべてのクラスタ設定作業への、完全な読み込み/書き込みのアクセス権を持ちます。ただし、[アクセス制御リスト\(ACL\)](#)を使用すれば、より詳細なアクセス権限を定義することができます。

CRMでACLが有効になっている場合、Hawk2で利用できる機能は、ユーザ役割と割り当てられたアクセスパーミッションごとに異なります。Hawk2の[履歴エクスプローラ]は、ユーザ `hacluster` のみが実行できます。

7.3.1 左のナビゲーションバー

[管理]

- [Status (状態)]: クラスタの現在の状態の概要が表示されます(`crmsh`の `crm status` と同様です)。詳細については、[7.8.1項「単一クラスタの監視」](#)を参照してください。クラスタに `guest nodes` (`pacemaker_remote` デーモンが実行されているノード)が含まれる場合、それらのノードも表示されます。この画面はほぼリアルタイムで更新され、ノードまたはリソースの状態に変化があった場合、ほとんど瞬時に表示されます。
- [ダッシュボード]: 複数のクラスタを監視できます(Geoクラスタがセットアップされている場合は、別のサイトにあるクラスタも監視できます)。詳細については、[7.8.2項「複数のクラスタの監視」](#)を参照してください。クラスタに `guest nodes` (`pacemaker_remote` デーモ

ンが実行されているノード)が含まれる場合、それらのノードも表示されます。この画面はほぼリアルタイムで更新され、ノードまたはリソースの状態に変化があった場合、ほとんど瞬時に表示されます。

- [履歴]: [履歴エクスプローラー]を開いてクラスタレポートを生成できます。詳細については、[7.10項「クラスタ履歴の表示」](#)を参照してください。

[環境設定]

- [リソースの追加]: リソース設定画を開きます。詳細については、[7.5項「クラスタリソースの設定」](#)を参照してください。
- [制約の追加]: 制約設定画面を開きます。詳細については、[7.6項「制約の設定」](#)を参照してください。
- [ウィザード]: さまざまなウィザードを選択できます。これにより、DRBDブロックデバイスなどの特定のワークロード用のリソースをウィザードに従って作成できます。詳細については、[7.5.2項「ウィザードを使用したリソースの追加」](#)を参照してください。
- [設定の編集]: リソース、制約、ノード名と属性、タグ、アラート、フェンシングトポロジなどを編集できます。
- [クラスタ設定]: グローバルクラスタオプション、およびリソースと操作のデフォルトを変更できます。詳細については、[7.4項「グローバルクラスタオプションの設定」](#)を参照してください。
- [Command Log (コマンドログ)]: Hawk2によって最近実行されたcrmshコマンドのリストを表示します。

[アクセス制御]

- [Roles (役割)]: アクセス制御リスト(CIBへのアクセス権を記述した一連のルール)に対して役割を作成できる画面を開きます。詳細については、[手順12.2「Hawk2によるMonitor役割の追加」](#)を参照してください。
- [Targets (ターゲット)]: アクセス制御リストのターゲット(システムユーザ)を作成して、そのターゲットに役割を割り当てることができる画面を開きます。詳細については、[手順12.3「Hawk2によるターゲットの役割割当」](#)を参照してください。

7.3.2 最上位の行

Hawk2の最上位の行には、次のエントリが表示されます。

- [バッチ]: クリックすると、バッチモードに切り替わります。これにより、変更をシミュレートしてステージングし、それらの変更を単一のトランザクションとして適用できます。詳細については、[7.9 項「バッチモードの使用」](#)を参照してください。
- [ユーザ名]: Hawk2用の設定ができます(たとえば、Webインタフェースの言語の設定やSTONITHを無効にした場合に警告を表示するかなどの設定)。
- [Help (ヘルプ)]: SUSE Linux Enterprise High Availability Extensionドキュメントにアクセスしたり、リリースノートを参照したり、バグを報告したりします。
- [ログアウト]: クリックするとログアウトします。

7.4 グローバルクラスタオプションの設定

グローバルクラスタオプションは、一定の状況下でのクラスタの動作を制御します。これらは、セットにグループ化され、Hawk2、crmshなどのクラスタ管理ツールで表示し、変更することができます。事前に定義されている値は、通常は、そのまま保持できます。ただし、クラスタの主要機能を正しく機能させるには、クラスタの基本的なセットアップ後に、次のパラメータを調整する必要があります。

- グローバルオプションno-quorum-policy
- グローバルオプションstonith-enabled

手順 7.3: グローバルクラスタオプションを変更する

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーから、[クラスタ設定]を選択します。
[クラスタ設定]画面が開きます。グローバルクラスタオプションとその現在の値が表示されます。
画面の右側にパラメータの簡単な説明を表示するには、マウスポインタをパラメータに合わせます。

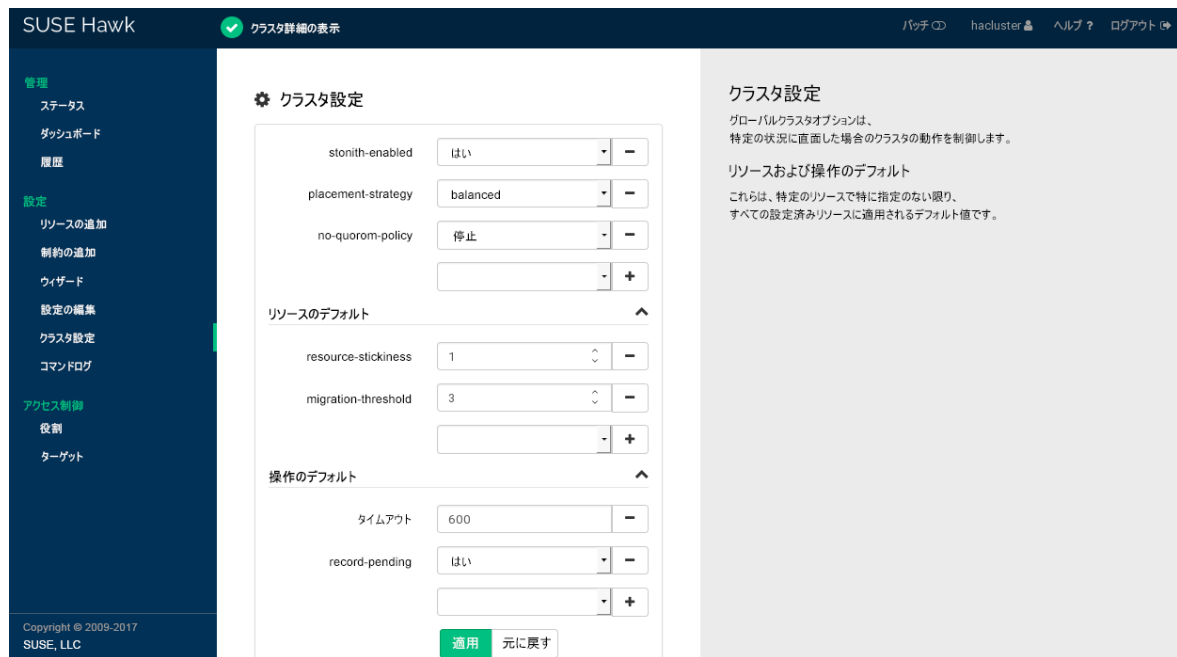


図 7.1: HAWK2 - クラスタの設定

3. [no-quorum-policy]および[stonith-enabled]の値を確認し、必要に応じて調整します。
 - a. [no-quorum-policy]を適切な値に設定します。詳細については、6.2.2項「グローバルオプションno-quorum-policy」を参照してください。
 - b. 何らかの理由でフェンシングを無効にする必要がある場合は、[stonith-enabled]をnoに設定します。通常のクラスタ操作にはSTONITHデバイスの使用が必要なため、デフォルトでは、trueに設定されています。デフォルト値では、クラスタは、STONITHリソースが設定されていない場合にはリソースの開始を拒否します。

❗ 重要: STONITHがない場合はサポートなし

- クラスタにはノードフェンシングメカニズムが必要です。
 - グローバルクラスタオプション stonith-enabled および startup-fencing を true に設定する必要があります。これらを変更するとサポートされなくなります。
- c. クラスタ設定からパラメータを削除するには、パラメータの横の[マイナス]アイコンをクリックします。パラメータを削除すると、クラスタはそのパラメータがデフォルト値に設定されている場合と同様に動作します。

- d. クラスタ設定に新たなパラメータを追加するには、ドロップダウンボックスから選択します。
4. [リソースのデフォルト]または[操作のデフォルト]を変更する必要がある場合は、次のような処理を実行します。
- a. 値を調整するには、ドロップダウンボックスから別の値を選択するか、値を直接編集します。
 - b. 新しいリソースのデフォルトまたは操作のデフォルトを追加するには、空のドロップダウンボックスから1つ選択し、値を入力します。デフォルト値が存在する場合は、Hawk2から自動的に提示されます。
 - c. パラメータを削除するには、その横の[マイナス]アイコンをクリックします。[リソースのデフォルト]と[操作のデフォルト]に値が指定されていない場合、クラスタは6.3.6項「リソースオプション(メタ属性)」および6.3.8項「リソース操作」にドキュメントされているデフォルト値を使用します。
5. 変更内容を確認します。

7.5 クラスタリソースの設定

クラスタの管理者は、クラスタ内のサーバ上の各リソースや、サーバ上で実行する各アプリケーションに対してクラスタリソースを作成する必要があります。クラスタリソースには、Webサイト、メールサーバ、データベース、ファイルシステム、仮想マシン、およびユーザが常時使用できるその他のサーバベースのアプリケーションまたはサービスなどが含まれます。

作成できるリソースタイプの概要については、6.3.3項「リソースのタイプ」を参照してください。リソースの基本情報(ID、クラス、プロバイダ、およびタイプ)を指定すると、Hawk2によって次のカテゴリが表示されます。

パラメータ(インスタンス属性)

リソースが制御するサービスのインスタンスを決定します。詳細については、6.3.7項「インスタンス属性(パラメータ)」を参照してください。

リソースを作成する際、Hawk2は必要なパラメータを自動的に表示します。これらを編集して、有効なリソースの設定を取得します。

操作

リソース監視に必要です。詳細については、6.3.8項「リソース操作」を参照してください。

リソースを作成する際、Hawk2は、重要なリソース操作を表示します(monitor、start、および stop)。

メタ属性

特定のリソースの処理方法をCRMに指示します。詳細については、[6.3.6項「リソースオプション \(メタ属性\)」](#)を参照してください。

リソースを作成する際、Hawk2はそのリソースの重要なメタ属性を自動的にリストにします(たとえばリソースの初期状態を定義する `target-role` 属性です。デフォルトでは `Stopped` に設定されているため、リソースはすぐには始動しません)。

使用率

特定のリソースがノードから要求する容量をCRMに指示します。詳細については、[7.6.8項「負荷インパクトに基づくリソース配置の設定」](#)を参照してください。

これらのカテゴリのエントリと値は、リソースの作成中に調整することも、後から調整することもできます。

7.5.1 現在のクラスタ設定の表示(CIB)

クラスタ管理者はクラスタ設定を知る必要がある場合があります。Hawk2は、現在の設定をcrmシェル構文で、XMLとして、およびグラフとして表示できます。クラスタ設定をcrmシェル構文で表示するには、左ナビゲーションバーから、[Edit Configuration (設定の編集)]を選択し、[Show (表示)]をクリックします。代わりに設定をraw XMLで表示するには、[XML]をクリックします。CIBで設定されたノードとリソースのグラフィカルな表現を示すには、[グラフ]をクリックします。リソース間の関係も表示されます。

7.5.2 ウィザードを使用したリソースの追加

Hawk2ウィザードは、仮想IPアドレスやSDB STONITHリソースなどの単純なリソースを設定する場合に便利です。また、DRBDブロックデバイスやApache Webサーバのリソース設定などの、複数リソースを含む複雑な設定においても役立ちます。設定手順をウィザードに従って進めることができ、入力が必要なパラメータについては情報が提供されます。

手順 7.4: リソースウィザードの使用

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーから、[ウィザード]を選択します。
3. ウィザードの横にある下矢印アイコンをクリックして個々のカテゴリを展開し、目的のウィザードを選択します。

4. 画面の指示に従います。最後の設定手順が完了したら、[Verify (検証)]を選択して、入力した値を検証します。
Hawk2が実行するアクションと、設定の内容が表示されます。設定によっては、[適用]を選択して設定を適用する前に、root パスワードの入力を求めるプロンプトが表示されることがあります。

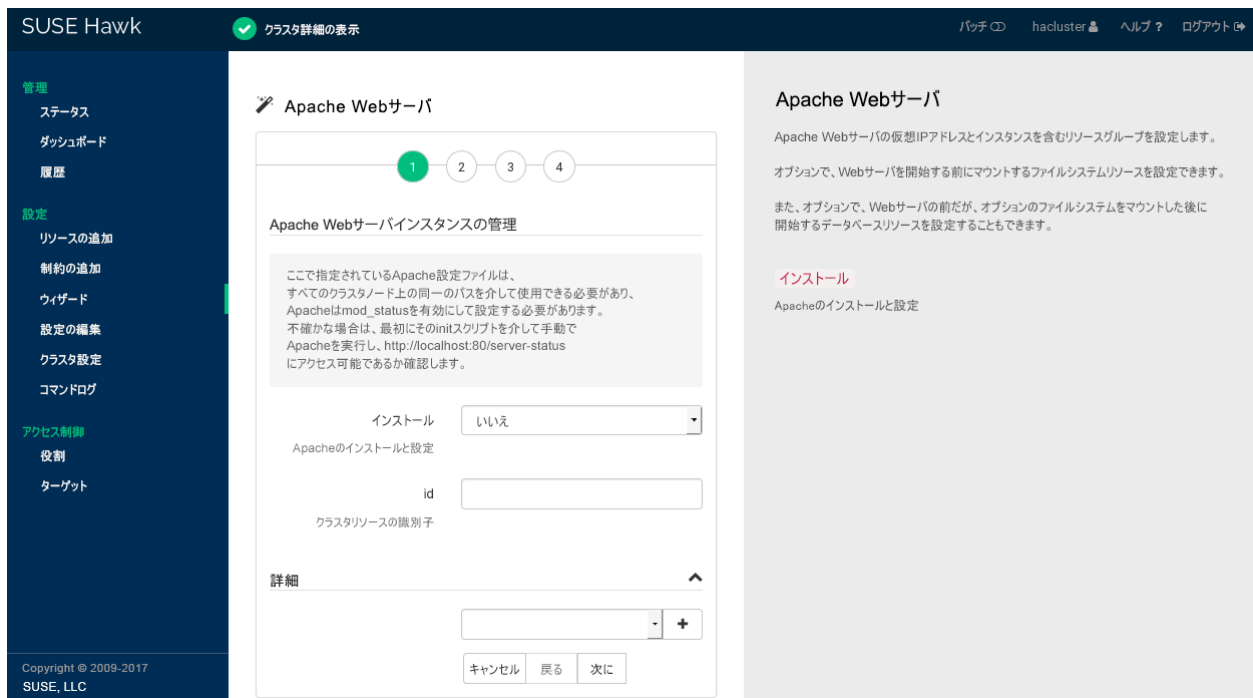


図 7.2: HAWK2 - APACHE WEBサーバ用のウィザード

7.5.3 単純なリソースの追加

最も基本的なタイプのリソースを作成するには、次の手順に従います。

手順 7.5: プリミティブリソースの追加

1. Hawk2にログインします。

`https://HAWKSERVER:7630/`

2. 左のナビゲーションバーから、[Add Resource (リソースの追加)] > [Primitive (プリミティブ)]の順に選択します。
3. 固有の[リソースID]を入力します。
4. リソース設定の基にするリソーステンプレートが存在する場合は、[テンプレート]で目的のテンプレートを選択します。テンプレートの設定の詳細については、[手順7.6「リソーステンプレートの追加」](#)を参照してください。

5. [クラス]で、使用するリソースエージェントのクラスを選択します。lsb、ocf、service、stonith、または systemd から選択できます。詳細については、6.3.2項「サポートされるリソースエージェントクラス」を参照してください。
6. ocf をクラスとして選択した場合、OCFリソースエージェントの[プロバイダ]を指定します。OCFの指定によって、複数のベンダが同じリソースエージェントを提供できるようになります。
7. [タイプ]リストから、使用するリソースエージェントを選択します(たとえば[IPaddr]または[Filesystem])。このリソースエージェントの簡単な説明が表示されます。これで、リソースの基本情報が指定されました。



注記

[タイプ]リストに表示される選択肢は、選択した[クラス](OCFリソースの場合は、[プロバイダ]も)によって異なります。

The screenshot shows the 'Primitive Creation' page in the SUSE Hawk interface. The main form has the following fields:

- Resource ID:
- Template:
- Class:
- Provider:
- Type:

Below these fields are four expandable sections, each with a downward arrow icon:

- パラメータ (Parameters)
- 操作 (Actions)
- メタ属性 (Meta-attributes)
- 利用率 (Usage)

At the bottom of the form are two buttons: '作成' (Create) and '戻る' (Back).

The right sidebar contains the following text:

プリミティブ
プリミティブリソースは、リソースの最も基本的なタイプです。
プリミティブを作成するには、IDを定義し、クラス、(プロバイダ)、タイプなどのいくつかのパラメータを指定します。
プリミティブリソースの一例は `ocf:heartbeat:IPaddr` で、クラスタのフローティングIPアドレスの設定に使用できます。

NFSサーバの管理
Nfsserverは、Linux nfsサーバがLinux-HAのフェールオーバー対応リソースとして管理するのに役立ちます。Linux固有のNFSの実装詳細に依存するため、他のプラットフォームに移植可能ではないとみなされます。

クラス
リソースエージェントが準拠する標準

図 7.3: HAWK2 - プリミティブリソース

8. Hawk2によって提案された[パラメータ]、[操作]、および[メタ属性]を保持する場合は、[作成]をクリックして設定を終了します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。
パラメータ、操作、またはメタ属性を調整するには、7.5.5項「リソースの変更」を参照してください。リソースの[利用率]属性を設定するには、手順7.21「リソースが要求する容量の設定」を参照してください。

7.5.4 リソーステンプレートの追加

類似した設定のリソースを多く作成する最も簡単な方法は、リソーステンプレートを定義することです。リソーステンプレートを定義した後は、プリミティブの中や、特定のタイプの制約で参照できるようになります。リソーステンプレートの機能と使用方法の詳細については、6.5.3項「リソーステンプレートと制約」を参照してください。

手順 7.6: リソーステンプレートの追加

リソーステンプレートは、プリミティブリソースと同様の方法で設定します。

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーから、[Add Resource(リソースの追加)] > [Template (テンプレート)]の順に選択します。
3. 固有の[リソースID]を入力します。
4. 手順7.5「プリミティブリソースの追加」のステップ 5以降の手順に従います。

7.5.5 リソースの変更

リソースの作成後、必要に応じてパラメータ、操作、またはメタ属性を調整することで、いつでもその設定を編集できます。

手順 7.7: パラメータ、操作、またはメタ属性の変更

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. Hawk2の[状態]画面で、[Resources (リソース)]リストに移動します。
3. [操作]列で、変更したいリソースまたはグループの横にある下矢印アイコンをクリックして[編集]を選択します。
リソース設定画面が開きます。

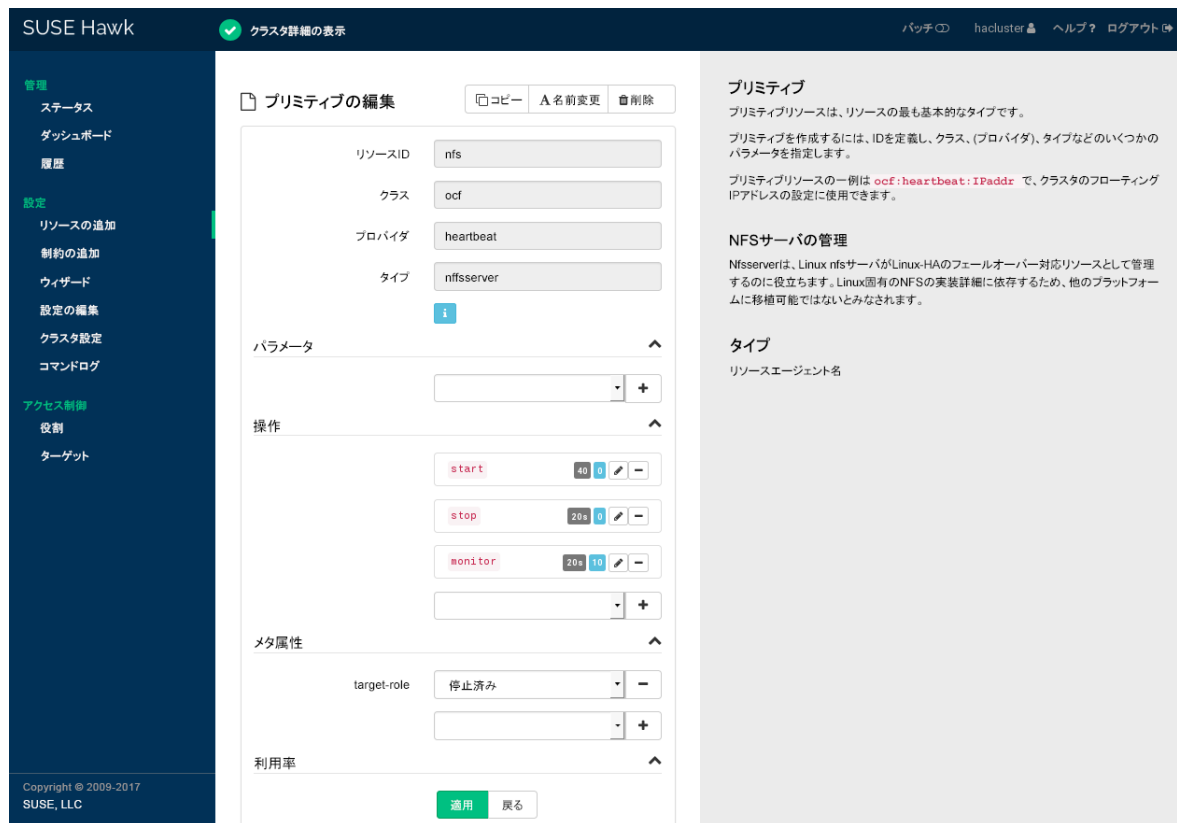


図 7.4: HAWK2 - プリミティブリソースの編集

4. 新たなパラメータ、操作、またはメタ属性を追加するには、空のドロップダウンボックスから項目を選択します。
5. [操作]カテゴリの値を編集するには、それぞれのカテゴリの[編集]アイコンをクリックして操作の別の値を入力し、[適用]をクリックします。
6. 完了したら、リソース設定画面で[適用] ボタンをクリックして、パラメータ、操作、またはメタ属性の変更内容を確認します。
画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。

7.5.6 STONITHリソースの追加

！ 重要: STONITHがない場合はサポートなし

- クラスタにはノードフェンシングメカニズムが必要です。
- グローバルクラスタオプション `stonith-enabled` および `startup-fencing` を `true` に設定する必要があります。これらを変更するとサポートされなくなります。

デフォルトでは、グローバルクラスタオプション `stonith-enabled` は `true` に設定されています。STONITHリソースが定義されていない場合、クラスタはどのリソースの開始も拒否します。1つ以上のSTONITHリソースを設定して、STONITHのセットアップを完了します。SBD、libvirt (KVM/Xen)、またはvCenter/ESX ServerのSTONITHリソースを追加する場合、Hawk2ウィザードを使用するのが最も簡単な方法です(7.5.2項「ウィザードを使用したリソースの追加」を参照してください)。STONITHは他のリソースと同様に設定しますが、その動作はいくつかの点で異なります。詳細については、10.3項「STONITHのリソースと環境設定」を参照してください。

手順 7.8: STONITHリソースの追加

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーから、[Add Resource (リソースの追加)] > [Primitive (プリミティブ)]の順に選択します。
3. 固有の[リソースID]を入力します。
4. [クラス]リストで、リソースエージェントクラスとして[stonith]を選択します。
5. [タイプ]リストから、使用しているSTONITHデバイスを制御するためのSTONITHプラグインを選択します。このプラグインの簡単な説明が下に表示されます。
6. Hawk2は、自動的にそのリソースに必要な[パラメータ]を表示します。それぞれのパラメータの値を入力します。
7. Hawk2は、重要なリソース[操作]を表示し、デフォルト値を提案します。ここで設定を変更しない場合、確定するとすぐに、Hawk2は提案した操作およびデフォルト値を追加します。
8. 変更理由がない場合は、デフォルトの[メタ属性]設定を保持します。

The screenshot shows the SUSE Hawk web interface for creating a primitive resource. The main form is titled 'プリミティブの作成' (Primitive Creation). It contains several sections: 'リソースID' (Resource ID) with 'stonith-1', 'テンプレート' (Template), 'クラス' (Class) with 'stonith', 'プロバイダ' (Provider) with 'apcmaster', and 'タイプ' (Type) with 'apcmaster'. The 'パラメータ' (Parameters) section includes 'ipaddr' (10.161.15.43), 'ログイン' (login: stonithmaster), and 'パスワード' (password: STRONG_PASSWORD). The '操作' (Actions) section has buttons for 'start', 'stop', and 'monitor'. The 'メタ属性' (Meta-properties) section includes 'target-role' set to '停止' (stop). At the bottom, there are '作成' (Create) and '戻る' (Back) buttons. The right sidebar contains text about 'プリミティブ' (Primitive) and 'APC MasterSwitch'.

図 7.5: HAWK2 - STONITHリソース

- 変更を確認して、STONITHリソースを作成します。
画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。

フェンシングを設定するには、制約を追加します。詳細については、[第10章「フェンシングとSTONITH」](#)を参照してください。

7.5.7 クラスタリソースグループの追加

クラスタリソースの中には、他のコンポーネントやリソースに依存しているものもあります。このような場合、各コンポーネントまたはリソースは特定の順番で起動し、同じサーバ上で動作する必要があります。この設定を簡単にするため、SUSE Linux Enterprise High Availability Extensionは、グループのコンセプトをサポートしています。

リソースグループには、一緒の場所で見つけ、連続して開始し、逆の順序で停止する必要のあるリソースセットが含まれます。リソースグループの例と、グループとそのプロパティの詳細について、[6.3.5.1項「グループ」](#)を参照してください。



注記: 空のグループ

グループには1つ以上のリソースを含む必要があります。空の場合は設定は無効になります。グループの作成中に、Hawk2ではさらにプリミティブを作成し、それらをグループに追加できます。詳細については、[7.7.1項「リソースとグループの編集」](#)を参照してください。

手順 7.9: リソースグループを追加する

1. Hawk2にログインします。

`https://HAWKSERVER:7630/`

2. 左ナビゲーションバーから、[Add Resource (リソースの追加)] > [Group (グループ)]の順に選択します。
3. 固有の[グループID]を入力します。
4. グループメンバーを定義するには、[子]リストで1つまたは複数のエントリを選択します。グループメンバーを再ソートするには、右側の「ハンドル」アイコンを使用して、メンバーを目的の順序にドラッグアンドドロップします。
5. 必要に応じて、[メタ属性]を変更または追加します。
6. [作成]をクリックして、設定を完了します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。

図 7.6: HAWK2 - リソースグループ

7.5.8 クローンリソースの追加

特定のリソースをクラスタ内の複数のノードで同時に実行することが必要な場合には、それらのリソースをクローンとして設定します。クローンとして設定できるリソースの一例は、OCFS2などのクラスタファイルシステムの `ocf:pacemaker:controld` です。標準のリソースまたはリソースグループであれば、どれでもクローンを作成できます。クローンリソースの各インスタンスは、すべて同じ動作にすることができます。ただし、ホストされているノードに応じて設定を変えることもできます。

利用可能なリソースクローンのタイプの概要は、[6.3.5.2項「クローン」](#)を参照してください。



注記: クローンの子リソース

クローンには、プリミティブまたはグループのいずれかを子リソースとして含めることができます。Hawk2では、クローンの作成中に子リソースを作成したり変更したりすることはできません。クローンを追加する前に、子リソースを作成し、必要に応じて設定しておいてください。詳細については、[7.5.3項「単純なリソースの追加」](#)または[7.5.7項「クラスタリソースグループの追加」](#)を参照してください。

手順 7.10: クローンリソースの追加

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーから、[Add Resource (リソースの追加)] > [Clone (クローン)]の順に選択します。
3. 固有の[クローンID]を入力します。
4. [子リソース]リストから、クローンのサブリソースとして使用するプリミティブまたはグループを選択します。
5. 必要に応じて、[メタ属性]を変更または追加します。
6. [作成]をクリックして、設定を完了します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。

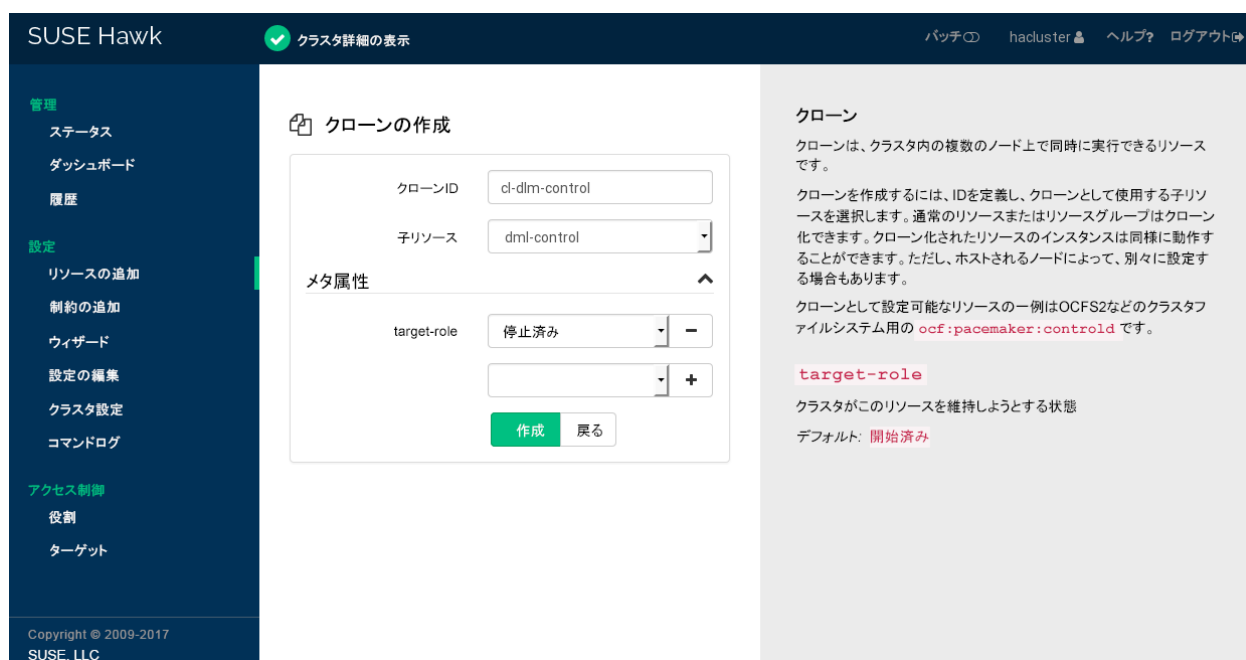


図 7.7: HAWK2 - クローンリソース

7.5.9 マルチステートリソースの追加

マルチステートリソースは、クローンが得意とするところです。これにより、インスタンスを2つの動作モード(`active/passive`、`primary/secondary`、または`master/slave`と呼ばれます)のいずれかに設定できます。マルチステートリソースは、グループまたは通常リソースを1つだけ含む必要があります。リソースの監視または制約を設定する場合、マルチステートリソースには、単純なリソースとは異なる要件があります。詳細については、『Pacemaker Explained』(<http://www.clusterlabs.org/doc/> から入手可)を参照してください。特に、「Multi-state - Resources That Have Multiple Modes」のセクションを参照してください。



注記: マルチステートリソースの子リソース

マルチステートリソースには、プリミティブまたはグループのいずれかを子リソースとして含めることができます。Hawk2では、マルチステートリソースの作成中に子リソースを作成したり変更したりすることはできません。マルチステートリソースを追加する前に、子リソースを作成し、必要に応じて設定しておいてください。詳細については、7.5.3項「単純なリソースの追加」または7.5.7項「クラスタリソースグループの追加」を参照してください。

手順 7.11: マルチステートリソースの追加

1. Hawk2にログインします。

https://HAWKSERVER:7630/

2. 左のナビゲーションバーから、[Add Resource (リソースの追加)] > [Multi-state (マルチステート)]の順に選択します。
3. [マルチステートID]に固有のIDを入力します。
4. [子リソース]リストから、マルチステートリソースのサブリソースとして使用するプリミティブまたはグループを選択します。
5. 必要に応じて、[メタ属性]を変更または追加します。
6. [作成]をクリックして、設定を完了します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。



図 7.8: HAWK2 - マルチステートリソース

7.5.10 タグの使用によるリソースのグループ化

タグは、コロケーションの作成や関係の順序付けを行わずに、複数のリソースをただちに参照する方法です。タグを使用して、概念的に関連するリソースをグループ化できます。たとえば、データベースに関連する複数のリソースがある場合、関連するすべてのリソースを database という名前のタグに追加できます。

同じタグに属するすべてのリソースは、1つのコマンドで開始または停止できます。

1. Hawk2にログインします。

`https://HAWKSERVER:7630/`

2. 左のナビゲーションバーから、[Add Resource (リソースの追加)] > [Tag (タグ)]の順に選択します。
3. [タグID]に固有のIDを入力します。
4. [オブジェクト]リストから、このタグで参照するリソースを選択します。
5. [作成]をクリックして、設定を完了します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。



図 7.9: HAWK2 - タグ

7.5.11 リソース監視の設定

High Availability Extensionは、ノードの障害を検出するだけでなく、ノードの個々のリソースで障害が発生した場合、そのことも検出します。リソースが実行中であるかどうか確認するには、そのリソースにリソースの監視を設定します。通常、リソースは動作中にのみ、クラスタによって監視されます。しかし、同時実行違反を検出するために、停止されるリソースの監視も設定する必要があります。リソースを監視するには、タイムアウト、開始遅延のいずれか、または両方と、間隔を指定します。間隔の指定によって、CRMにリソースステータスの確認頻度を指示します。start または stop 操作に対する timeout など、特定のパラメータも設定できます。

1. Hawk2にログインします。

`https://HAWKSERVER:7630/`

2. 手順7.5「プリミティブリソースの追加」の説明に従ってリソースを追加するか、既存のプリミティブを選択して編集します。
Hawk2は、重要な[操作](start、stop、monitor)を自動的に表示し、デフォルト値を提案します。
提案された各値に属する属性を参照するには、マウスポインタをそれぞれの値に合わせます。

操作 ^

start	20	✎	-
stop	20	✎	-
monitor	20	10	✎ -

timeout

3. start または stop 操作に対して提案された timeout の値を変更するには、次の手順に従います。
 - a. 操作の隣のペンアイコンをクリックします。
 - b. 表示されるダイアログで、timeout パラメータに別の値(例: 10)を入力し、変更内容を確認します。
4. monitor 操作に対して提案された [interval] の値を変更するには、次の手順に従います。
 - a. 操作の隣のペンアイコンをクリックします。
 - b. 表示されるダイアログで、interval に対して監視間隔の別の値を入力します。
 - c. リソースが停止されている場合にリソース監視を設定するには、次の手順に従います。
 - i. 下にある空のドロップダウンボックスから 役割 項目を選択します。
 - ii. [役割] ドロップダウンボックスから、[Stopped (停止済み)] を選択します。
 - iii. [適用] をクリックし、変更内容を確認して操作のダイアログを閉じます。
5. リソース設定画面で変更内容を確認します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。

リソースモニタが障害を検出した場合の処理については、[6.4項「リソース監視」](#)を参照してください。

リソースの障害を表示するには、Hawk2で[状態]画面に切り替えて、関係するリソースを選択します。[操作]列で、下矢印アイコンをクリックして[最近のイベント]を選択します。表示されるダイアログに、リソースに対して実行された最近のアクションが一覧表示されます。障害は赤で表示されます。リソースの詳細を表示するには、[操作]列で虫眼鏡アイコンをクリックします。

Q nfs

プリミティブ

エージェント

ocf:heartbeat:nfsserver

メタ属性

▼

target-role

停止済み

操作

▼

名前	タイムアウト	間隔
開始	40	0
停止	20s	0
モニタ	20s	10

制約

▼

ID	タイプ	スコア	宛先
----	-----	-----	----

閉じる

図 7.10: HAWK2 - リソース詳細

7.6 制約の設定

すべてのリソースを設定したら、クラスタがそれらを扱う方法を指定します。リソース制約を使えば、リソースがどのクラスタノードで実行されるか、リソースをどの順番でロードするか、そして特定のリソース型のどのリソースに依存するかを指定することができます。

利用可能な制約のタイプの概要は、[6.5.1項「制約のタイプ」](#)を参照してください。制約を定義する際には、スコアも指定する必要があります。スコアおよびクラスタでのそれらの意味の詳細については、[6.5.2項「スコアと無限大」](#)を参照してください。

7.6.1 場所制約の追加

場所制約は、リソースを実行できるノード、実行に適したノード、または実行できないノードを決定します。たとえば、場所制約により、特定のデータベースに関連するすべてのリソースを同じノードに配置します。

手順 7.14: 場所制約の追加

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーから、[Add Constraint (制約の追加)] > [Location (場所)]の順に選択します。
3. 固有の[制約ID]を入力します。
4. [Resources (リソース)]のリストから、制約を定義するリソースを1つまたは複数選択します。
5. [スコア]にスコアを入力します。スコアはこのリソース制約に割り当てる値を示します。正の値は、次のステップで指定する[ノード]でリソースを実行できることを示します。負の値は、リソースをそのノードで実行すべきではないことを示します。スコアの高い制約は、それよりもスコアが低い制約より先に適用されます。
使用頻度の高い次の値は、ドロップダウンボックスからも設定できます。
 - ノードで強制的にリソースを実行するには、矢印アイコンをクリックして 常に を選択します。これにより、スコアは INFINITY に設定されます。
 - ノードでリソースを実行しない場合、矢印アイコンをクリックして Never (実行しない) を選択します。これにより、スコアは -INFINITY に設定され、リソースはそのノードで実行してはならないことになります。
 - スコアを 0 に設定するには、矢印アイコンをクリックして Advisory (推奨値) を選択します。これにより制約が無効になります。これは、リソース検出を設定してもリソースを制約したくない場合に便利です。
6. [ノード]を選択します。
7. [作成]をクリックして、設定を完了します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。

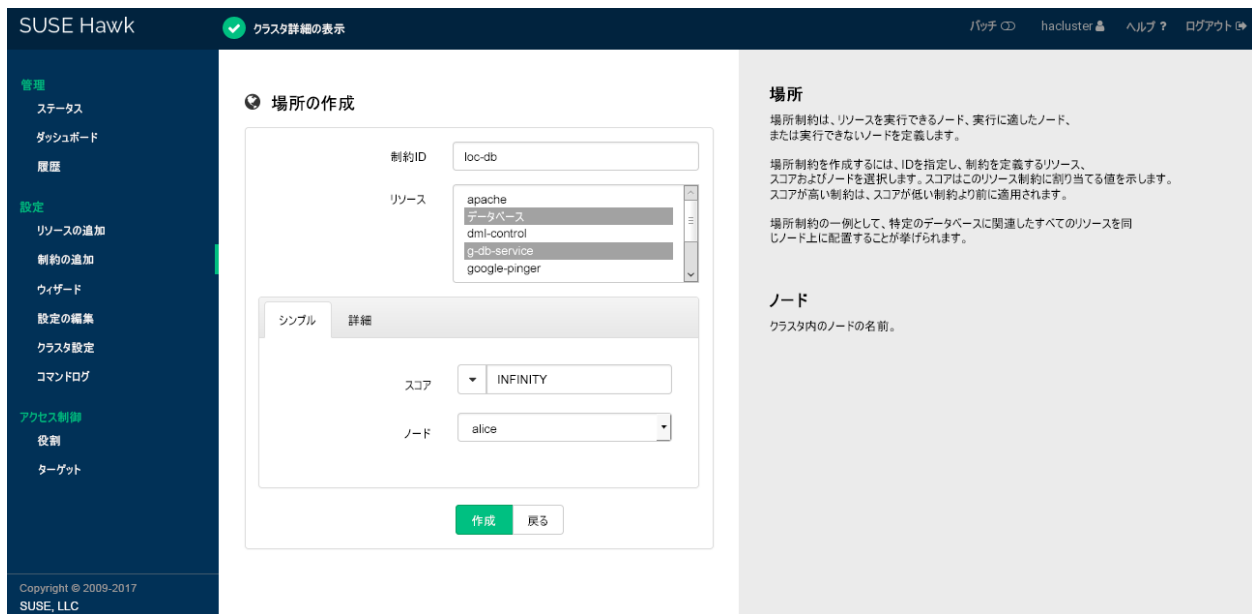


図 7.11: HAWK2 - 場所制約

7.6.2 コロケーション制約の追加

コロケーション制約は、ノード上で一緒に実行可能な、または一緒に実行することが禁止されているリソースをクラスタに伝えます。コロケーション制約はリソース間の依存関係を定義するため、コロケーション制約を作成するには、少なくとも2つのリソースが必要です。

手順 7.15: コロケーション制約の追加

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーから、[Add Constraint (制約の追加)] > [Colocation (コロケーション)]の順に選択します。
3. 固有の[制約ID]を入力します。
4. [スコア]にスコアを入力します。スコアは複数のリソースの場所の関係を決定します。正の値は、リソースを同じノードで実行しなければならないことを示します。負の値は、リソースを同じノードで実行するべきではないことを示します。スコアと他の要因との組み合わせによって、ノードの配置先が決定します。
使用頻度の高い次の値は、ドロップダウンボックスからも設定できます。

- リソースを強制的に同じノードで実行する場合、矢印アイコンをクリックして Always (常に実行する) を選択します。これにより、スコアは INFINITY に設定されます。
 - リソースを同じノードで実行しない場合、矢印アイコンをクリックして Never (実行しない) を選択します。これにより、スコアは -INFINITY に設定され、リソースは同じノードで実行してはならないことになります。
5. 制約のリソースを定義するには、次の手順に従います。
 - a. [リソース] カテゴリのドロップダウンボックスから、リソース(またはテンプレート)を選択します。
リソースが追加され、下に新しい空のドロップダウンボックスが表示されます。
 - b. この手順を繰り返してリソースを追加します。
最上位のリソースは次のリソースなど順に依存するため、クラスタはまず最後のリソースを置く場所を決め、次にその決定に基づいて依存するものを配置していきます。制約が満たされないと、クラスタは依存するリソースが実行しないようにすることがあります。
 - c. コロケーション制約内のリソースの順序を入れ替えるには、リソースの横にある上矢印アイコンをクリックして、その上のエントリと入れ替えます。
 6. 必要に応じて、各リソースの他のパラメータ(Started (開始)、Stopped (停止)、Master (マスタ)、Slave (スレーブ)、Promote (昇格)、Demote (降格) など)を指定します。リソースの横にある空のドロップダウンリストをクリックして、目的のエントリを選択します。
 7. [作成]をクリックして、設定を完了します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。

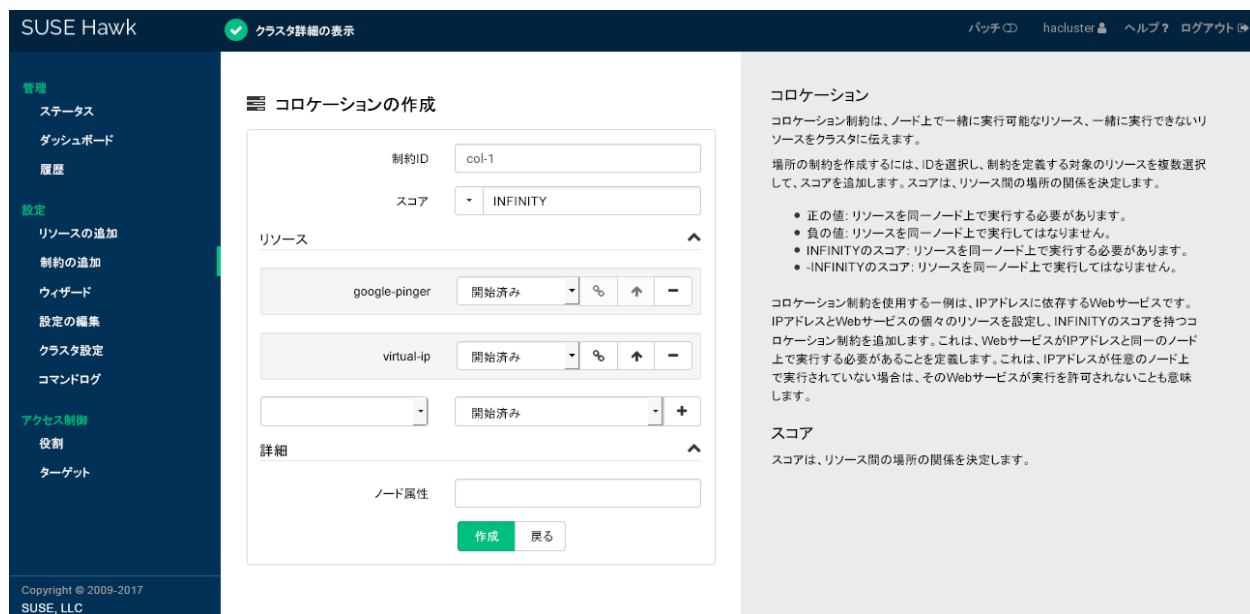


図 7.12: HAWK2 - コロケーション制約

7.6.3 順序制約の追加

順序制約は、リソースを開始および停止する順序を定義します。順序制約はリソース間の依存関係を定義するため、順序制約を作成するには、少なくとも2つのリソースを作成する必要があります。

手順 7.16: 順序制約の追加

1. Hawk2にログインします。

`https://HAWKSERVER:7630/`

2. 左のナビゲーションバーから、[Add Constraint (制約の追加)] > [Order (順序)]の順に選択します。
3. 固有の[制約ID]を入力します。
4. [スコア]にスコアを入力します。スコアがゼロより大きい場合、順序制約は必須になりますが、そうでない場合はオプションです。
使用頻度の高い次の値は、ドロップダウンボックスからも設定できます。

- 順序制約を必須にする場合、矢印アイコンをクリックして Mandatory (必須) を選択します。
 - 順序制約を提案にとどめる場合は、矢印アイコンをクリックして Optional (オプション) を選択します。
 - Serialize (順番に処理): リソースに対して2つの停止/開始アクションが同時に実行されないようにするには、矢印アイコンをクリックして Serialize (順番に処理) を選択します。これにより、1つのリソースの開始が完了しないと他のリソースを開始できなくなります。通常は、起動時にホストに高い負荷をかけるリソースに使用します。
5. 順序の制約の場合、オプション[シンメトリック]は常に有効にしてください。これは、リソースを停止するときには逆順で行うという指定です。
 6. 制約のリソースを定義するには、次の手順に従います。
 - a. [リソース]カテゴリのドロップダウンボックスから、リソース(またはテンプレート)を選択します。
リソースが追加され、下に新しい空のドロップダウンボックスが表示されます。
 - b. この手順を繰り返してリソースを追加します。
リソースは、最初に最上位のリソース、次に2番目のリソースという順序で開始されます。通常、リソースを停止するときには逆順で行われます。
 - c. 順序制約内のリソースの順序を入れ替えるには、リソースの横にある上矢印アイコンをクリックして、その上のエントリと入れ替えます。
 7. 必要に応じて、各リソースの他のパラメータ(Started (開始)、Stopped (停止)、Master (マスター)、Slave (スレーブ)、Promote (昇格)、Demote (降格)など)を指定します。リソースの横にある空のドロップダウンリストをクリックして、目的のエントリを選択します。
 8. 変更を確認して、設定を完了します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。



図 7.13: HAWK2 - 順序制約

7.6.4 制約のためにリソースセットを使用する

制約を定義するための別のフォーマットとして、**リソースセット**を使用することができます。これらは、**グループ**と同じ意味論に従います。

手順 7.17: 制約のためにリソースセットを使用する

1. 場所制約内でリソースセットを使用するには:
 - a. **手順7.14「場所制約の追加」**の説明に従って操作を進めます。ただし、**ステップ 4**は除きます。1つのリソースを選択する代わりに、**Ctrl** または **Shift** を押しながらマウスをクリックして、複数のリソースを選択します。これにより、場所制約内でリソースセットが作成されます。
 - b. 場所制約からリソースを削除するには、**Ctrl** を押しながらリソースを再度クリックして、選択解除します。
2. コロケーションまたは順序の制約内でリソースセットを使用するには:
 - a. **手順7.15「コロケーション制約の追加」**または**手順7.16「順序制約の追加」**の説明に従います。ただし、制約に対してリソースを定義する手順(**ステップ 5.a**または**ステップ 6.a**)は除きます。
 - b. 複数のリソースを追加します。

- c. リソースセットを作成するため、リソースの横にあるチェーンアイコンをクリックして、そのリソースを上のリソースにリンクします。リソースセットは、セットに属しているリソースの周囲のフレームによって示されます。
- d. リソースセット内の複数のリソースを結合したり、複数のリソースセットを作成したりできます。

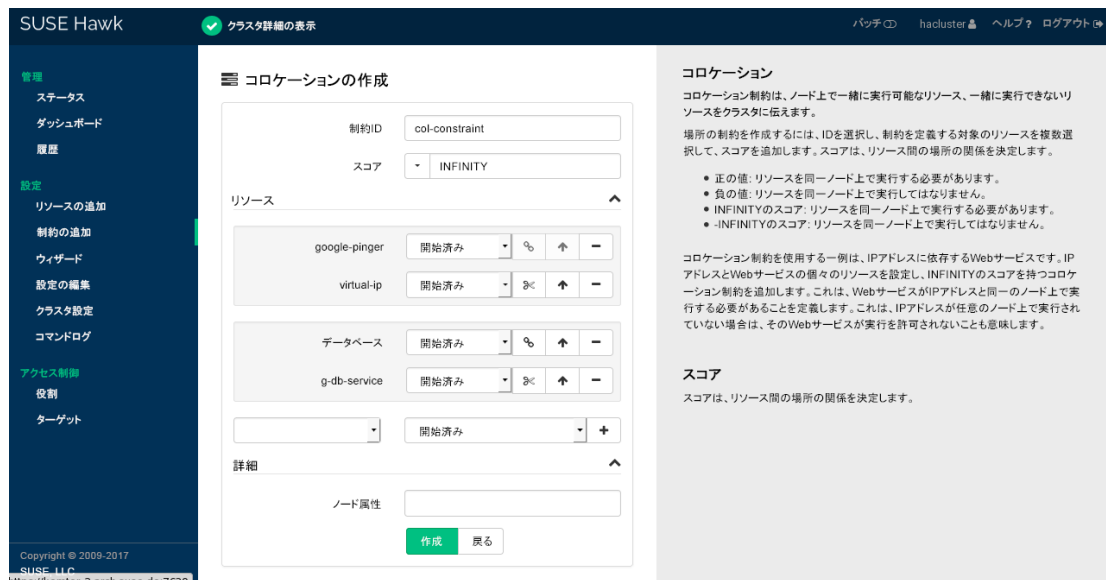


図 7.14: HAWK2 - コロケーション制約の2つのリソースセット

- e. 上のリソースからリソースをリンク解除するには、そのリソースの横にあるハサミアイコンをクリックします。

3. 変更を確認して、制約の設定を完了します。

7.6.5 その他の情報

制約の設定の詳細や、順序およびコロケーションの基本的な概念についての詳細なバックグラウンド情報は、<http://www.clusterlabs.org/doc/> で提供されているドキュメントを参照してください。

- 『Pacemaker Explained』の「Resource Constraints」の章
- 『Colocation Explained』
- 『オーダーの概要』

7.6.6 リソースフェールオーバーノードの指定

リソースに障害が発生すると、自動的に再起動されます。現在のノードで再起動できない場合、または現在のノードで N 回失敗した場合は、別のノードへのフェールオーバーが試行されます。新しいノードへのマイグレートを行う基準(`migration-threshold`)となるリソースの失敗をいくつか定義できます。クラスタ内に3つ以上ノードがある場合、特定のリソースのフェールオーバー先のノードはHigh Availabilityソフトウェアにより選択されます。

リソースがフェールオーバーする特定のノードを前もって指定するには、次の手順に従います。

手順 7.18: フェールオーバーノードの指定

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 手順 7.14「場所制約の追加」に記されている手順に従って、そのリソースの場所の制約を設定します。
3. 手順 7.7: パラメータ、操作、またはメタ属性の変更、ステップ 4に説明されている手順に従ってリソースに `migration-threshold` メタ属性を追加し、`migration-threshold` の値を入力します。INFINITYではない正の値を指定する必要があります。
4. リソースの失敗回数を自動的に失効させる場合は、手順 7.5: プリミティブリソースの追加、ステップ 4に説明されている手順に従って `failure-timeout` メタ属性をそのリソースに追加し、`failure-timeout` の [値] を入力します。

SUSE Hawk

クラスタ詳細の表示

バッチ 〇 hacluster ヘルプ? ログアウト

管理

- ステータス
- ダッシュボード
- 履歴
- 設定
- リソースの追加
- 制約の追加
- ウィザード
- 設定の編集
- クラスタ設定
- コマンドログ
- アクセス制御
- 役割
- ターゲット

Copyright © 2009-2017 SUSE, LLC

プリミティブの編集

コピー A 名前変更 削除

リソースID simple-testresource

クラス ocf

プロバイダ heartbeat

タイプ ダミー

パラメータ

操作

メタ属性

migration-threshold 1000

failure-timeout 10

利用率

適用 元に戻す 戻る

プリミティブを作成するには、IDを定義し、クラス、(プロバイダ)、タイプなどのいくつかのパラメータを指定します。

プリミティブリソースの一例は `ocf:heartbeat:IPaddr` で、クラスタのフローティングIPアドレスの設定に使用できます。

ステートレスリソースエージェントの例

これは、ダミーリソースエージェントです。実行中であるかどうかを追跡する以外は何も実行しません。その目的はテストを行うこととRAライターのテンプレートの役目を果たすることです。

NB: 下のアクションセクションで指定されたタイムアウトに注意してください。このタイムアウトは、エージェントが管理するリソースの種類にとって意味があるはずですが、最小タイムアウトであることが推奨されますが、すべての利用可能なリソースインスタンスをカバーできない場合があります。したがって、大きすぎる値にも、小さすぎる値にもせず、適切な値にしてください。最小タイムアウトは10秒を下回らないようにしてください。

タイプ

リソースエージェント名

5. リソースの初期設定として、追加のフェールオーバーノードを指定する場合は、追加の場所の制約を作成します。

マイグレーションしきい値と失敗カウントに関連したプロセスフローは、例6.8「マイグレーションしきい値 - プロセスフロー」に示されています。

リソースの失敗回数は、自動的に期限切れにする代わりに、いつでも、手動でクリーンアップすることもできます。詳細については、7.7.3項「リソースのクリーンアップ」を参照してください。

7.6.7 リソースフェールバックノードの指定(リソースの固着性)

ノードがオンライン状態に戻り、クラスタ内にある場合は、リソースが元のノードにフェールバックすることがあります。このことを防ぐ、またはリソースのフェールバック先として別のノードを指定するには、リソースの固着性の値を変更します。リソースの固着性は、リソースを作成するとき、または作成した後のどちらでも指定できます。

さまざまなリソース固着性値の意味については、6.5.5項「フェールバックノード」を参照してください。

手順 7.19: リソースの固着性を指定する

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 手順 7.7: パラメータ、操作、またはメタ属性の変更、ステップ 4に従って、resource-stickiness メタ属性をリソースに追加します。
3. resource-stickiness として、-INFINITYと INFINITY の間の値を指定します。



7.6.8 負荷インパクトに基づくリソース配置の設定

すべてのリソースが同等ではありません。Xenゲストなどの一部のリソースでは、そのホストであるノードがリソースの容量要件を満たす必要があります。配置したリソースの要件の合計が提供容量よりも大きくなった場合には、リソースのパフォーマンスが低下するか、または失敗します。

これを考慮に入れて、High Availability Extensionでは、次のパラメータを指定できます。

1. 一定のノードが提供する容量
2. 一定のリソースが要求する容量
3. リソースの配置に関する全体的なストラテジ

詳細と設定例については、6.5.6項「負荷インパクトに基づくリソースの配置」を参照してください。

使用属性は、リソースの要件と、ノードが提供する容量の両方を設定するために使用されます。リソースが要求する容量を設定するには、その前にノードの容量を設定する必要があります。

手順 7.20: ノードが提供する容量の設定

1. Hawk2にログインします。

`https://HAWKSERVER:7630/`

2. 左のナビゲーションバーで、[Status (状態)]を選択します。

3. [ノード]タブで、容量を設定するノードを選択します。
4. [操作]列で、下矢印アイコンをクリックして[編集]を選択します。
[ノードの編集]画面が開きます。
5. [使用率]の下で、使用属性の名前を空のドロップダウンボックスに入力します。
名前は任意です(RAM_in_GBなど)。
6. 属性を追加するには、[追加]アイコンをクリックします。
7. 属性の隣の空のテキストボックスに、属性値を入力します。値は整数にする必要があります。
8. 必要なだけ使用属性を追加し、これらの属性すべての値を追加します。
9. 変更内容を確認します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。

手順 7.21: リソースが要求する容量の設定

プリミティブリソースを作成するときや、既存のプリミティブリソースを編集するとき、特定のリソースがノードに要求する容量を設定します。

リソースに使用属性を追加する前に、[手順 7.20](#)で説明するように、クラスタノードの使用属性を設定しておく必要があります。

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 既存のリソースに使用属性を追加する場合: [7.7.1項「リソースとグループの編集」](#)に示すように、[Manage (管理)] > [状態]の順に移動して、[リソース設定]ダイアログを開きます。
新しいリソースを作成する場合: [7.5.3項「単純なリソースの追加」](#)に示すように、[Configuration (設定)] > [Add Resource (リソースの追加)]の順に移動して進みます。
3. [リソース設定]ダイアログで、[使用率]カテゴリに移動します。
4. 空のドロップダウンボックスから、[手順 7.20](#)でノードに対して設定した使用属性のいずれかを選択します。
5. 属性の隣の空のテキストボックスに、属性値を入力します。値は整数にする必要があります。
6. 必要なだけ使用属性を追加し、これらの属性すべての値を追加します。
7. 変更内容を確認します。画面上部に、アクションが成功したかどうかを示すメッセージが表示されます。

ノードが提供する容量とリソースが要求する容量を設定してから、配置ストラテジをグローバルクラスターオプションに設定します。そうしないと、容量設定は有効になりません。負荷のスケジュールに使用できるストラテジがいくつかあります。たとえば、負荷をできるだけ少ない数のノードに集中したり、使用可能なすべてのノードに均等に分散できます。詳細については、[6.5.6項「負荷インパクトに基づくリソースの配置」](#)を参照してください。

手順 7.22: 配置ストラテジを設定する

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーで、[クラスター設定]を選択し、各画面を開きます。グローバルクラスターオプション、およびリソースと操作のデフォルトが表示されます。
3. 画面上部にある空のドロップダウンボックスから、placement-strategyを選択します。
デフォルトでは、その値は[デフォルト]に設定され、使用属性と値が考慮されていないことを意味します。
4. 要件に応じて、[配置ストラテジ]を適切な値に設定します。
5. 変更内容を確認します。

7.7 クラスターリソースの管理

Hawk2では、クラスターリソースを設定することに加えて、[状態]画面で既存のリソースを管理することができます。この画面の概要については、[7.8.1項「単一クラスターの監視」](#)を参照してください。

7.7.1 リソースとグループの編集

既存のリソースを編集する必要がある場合は、[状態]画面に移動します。[操作]列で、変更したいリソースまたはグループの横にある下矢印アイコンをクリックして[編集]を選択します。

編集画面が表示されます。プリミティブリソースを編集する場合は、次の操作が使用できます。

プリミティブの操作

- リソースのコピー。
- リソースの名前変更 (そのIDの変更)。
- リソースの削除。

グループを編集する場合は、次の操作が使用できます。

グループの操作

- このグループに追加される新しいプリミティブの作成。
- グループの名前変更 (そのIDの変更)。
- グループメンバーを再ソートするには、右側の「ハンドル」アイコンを使用して、メンバーを目的の順序にドラッグアンドドロップします。

7.7.2 リソースの開始

クラスタリソースは、起動する前に、正しく設定されているようにします。たとえば、Apacheサーバをクラスタリソースとして使用する場合は、まず、Apacheサーバを設定します。クラスタでそれぞれのリソースを開始する前に、Apache設定を完了します。



注記: クラスタによって管理されるサービスには介入しないでください。

High Availability Extensionでリソースを管理しているときに、リソースを他の方法(クラスタ外で、たとえば、手動、ブート、再起動など)で開始したり、停止したりしてはなりません。High Availability Extensionソフトウェアが、すべてのサービスの開始または停止アクションを実行します。

ただし、サービスが適切に構成されているか確認したい場合は手動で開始しますが、High Availabilityが起動する前に停止してください。

現在クラスタで管理されているリソースへの介入については、まず、リソースを maintenance mode に設定します。詳細については、[手順16.5「リソースをHawk2を使用して保守モードにする」](#)を参照してください。

Hawk2でリソースを作成するときには、target-role メタ属性でその初期状態を設定することができます。その値を stopped に設定した場合、リソースは、作成後、自動的に開始することはありません。

手順 7.23: 新しいリソースを起動する

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーで、[Status (状態)]を選択します。[Resources (リソース)]のリストには[状態]も表示されます。

3. 開始するリソースを選択し、その[操作]列で[Start (開始)]アイコンをクリックします。継続するには、表示されるメッセージに対して確認します。
リソースが開始すると、Hawk2はすぐにリソースの[状態]を緑に変え、それがどのノードで実行されているかを表示します。

7.7.3 リソースのクリーンアップ

リソースは、失敗した場合は自動的に再起動しますが、失敗のたびにリソースの失敗回数が増加します。

リソースに対して `migration-threshold` が設定されていた場合、失敗回数が移行しきい値に達すると、そのリソースはそのノードでは実行されなくなります。

リソースの失敗回数は、(リソースに `failure-timeout` オプションを設定することにより)自動的にリセットするか、または次に示すように手動でリセットできます。

手順 7.24: リソースをクリーンアップする

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーで、[Status (状態)]を選択します。[Resources (リソース)]のリストには[Status (状態)]も表示されます。
3. クリーンアップするリソースに移動します。[操作]列で、下矢印ボタンをクリックして[Cleanup (クリーンアップ)]を選択します。継続するには、表示されるメッセージに対して確認します。
これにより、コマンド `crm resource cleanup` が実行され、すべてのノードでリソースがクリーンアップされます。

7.7.4 クラスタリソースの削除

リソースをクラスタから削除する必要がある場合は、次の手順に従って、設定エラーが発生しないようにします。

手順 7.25: クラスタリソースの削除

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 手順7.24「リソースをクリーンアップする」の説明に従って、すべてのノードでリソースをクリーンアップします。

3. リソースを停止します。

- a. 左のナビゲーションバーで、[Status (状態)]を選択します。[Resources (リソース)]のリストには[Status (状態)]も表示されます。
- b. [操作]列で、リソースの横にある[Stop (停止)]ボタンをクリックします。
- c. 継続するには、表示されるメッセージに対して確認します。
リソースが停止すると、[状態]列に変更が反映されます。

4. リソースを削除します。

- a. 左のナビゲーションバーで、[Edit Configuration (設定の編集)]を選択します。
- b. [Resources (リソース)]のリストで、それぞれのリソースに移動します。[操作]列で、リソースの横にある[Delete (削除)]アイコンをクリックします。
- c. 継続するには、表示されるメッセージに対して確認します。

7.7.5 クラスタリソースの移行

7.6.6項「リソースフェールオーバーノードの指定」で説明したように、ソフトウェアまたはハードウェアの障害時には、クラスタは定義可能な特定のパラメータ(たとえばマイグレーションしきい値やリソースの固着性など)に従って、リソースを自動的にフェールオーバー(マイグレート)させます。クラスタ内の別のノードにリソースを手動で移行させることもできます。または、リソースを現在のノードから出すかはユーザが判断し、どこに移動するかはクラスタに判断させます。

手順 7.26: リソースを手動で移行する

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーで、[Status (状態)]を選択します。[Resources (リソース)]のリストには[Status (状態)]も表示されます。
3. [Resources (リソース)]のリストで、それぞれのリソースを選択します。
4. [操作]列で、下矢印ボタンをクリックして[Migrate (移行)]を選択します。
5. 開くウィンドウには、次の選択肢があります。

- [Away from current node (現在のノードから離れる)]: これによって現在のノードに対して -INFINITY スコアによる場所の制約が作成されます。
- または、別のノードにリソースを移動できます。これによって移動先ノードに対して INFINITY スコアの場所の制約が作成されます。

6. 選択内容を確認します。

リソースを再び元に戻すには、次の手順に従います。

手順 7.27: リソースの移行解除

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーで、[Status (状態)]を選択します。[Resources (リソース)]のリストには[Status (状態)]も表示されます。
3. [Resources (リソース)]のリストで、それぞれのリソースに移動します。
4. [Operations (操作)]列で、下矢印ボタンをクリックして[Unmigrate (移行解除)]を選択します。継続するには、表示されるメッセージに対して確認します。
これによって、`crm_resource -U` コマンドが使用されます。リソースは元の場所に戻ることができます。あるいは現在の場所に残ることもあります(リソースの固着性によって)。

詳細については、<http://www.clusterlabs.org/doc/> から入手できる『Pacemaker Explained』を参照してください。特に、「Resource Migration」のセクションを参照してください。

7.8 クラスタの監視

Hawk2には、単一のクラスタおよび複数のクラスタを監視するための異なる画面があります: [状態]画面と[ダッシュボード]画面。

7.8.1 単一クラスタの監視

単一クラスタを監視するには、[状態]画面を使用します。Hawk2にログインした後で、[状態]画面がデフォルトで表示されます。右上隅のアイコンに、クラスタの状態の概要が表示されます。詳細情報を参照する場合は、次のカテゴリを確認します。

エラー

エラーが発生した場合、ページの上部に表示されます。

リソース

設定されているリソースが表示されます。その[状態]、[名前](ID)、[場所](リソースが実行されているノード)、リソースエージェントの[タイプ]が含まれます。[操作]列で、リソースを開始または停止するか、いくつかのアクションをトリガーするか、詳を表示できます。トリガ可能なアクションには、保守モードへのリソースの設定(または保守モードの削除)、リソースの別のノードへの移行、リソースのクリーンアップ、リソースイベントの表示、またはリソースの編集が含まれます。

ノード

ログイン先のクラスタサイトに属するノードが表示されます。ノードの[状態]および[名前]が含まれます。[Maintenance (保守)]および[Standby (スタンバイ)]列で、`maintenance`または`standby`フラグを設定または解除できます。[操作]列で、ノードの最新イベントや詳細情報を参照できます。たとえば、それぞれのノードに`utilization`、`standby`、または`maintenance`のどの属性が設定されているかを参照できます。

チケット

Geoクラスタリングでの使用向けにチケットを設定した場合にのみ表示されます。

ステータス	名前	場所	タイプ	操作
+	apache		ocf:heartbeat:Dummy	[Stop] [Refresh] [Search]
+	データベース	kemter-3	ocf:heartbeat:Dummy	[Stop] [Refresh] [Search]
+	dmi-control	kemter-3	ocf:heartbeat:Dummy	[Stop] [Refresh] [Search]
+	g-db-service		グループ(2)	[Stop] [Refresh] [Search]
+	google-pinger		ocf:heartbeat:Dummy	[Stop] [Refresh] [Search]
+	nfs		ocf:heartbeat:nfsserver	[Stop] [Refresh] [Search]
+	stonith		stonith:external/ipmi	[Stop] [Refresh] [Search]
+	virtual-ip		ocf:heartbeat:Dummy	[Stop] [Refresh] [Search]

図 7.15: HAWK2 - クラスタの状態

7.8.2 複数のクラスタの監視

複数のクラスタを監視するには、Hawk2 [Dashboard (ダッシュボード)]を使用します。[ダッシュボード]画面に表示されるクラスタ情報は、サーバ側に保管されています。これらは、クラスタノード間で同期が取られています(クラスタノード間にパスワード不要のSSHアクセスが設定されている場合)。詳細については、[D.2項「パスワード不要のSSHアカウントの設定」](#)を参照してください。ただし、Hawk2を実行するマシンは、その目的のためにクラスタの一部である必要はなく、別個の無関係のシステムで構いません。

Hawk2で複数のクラスタを監視するには、一般的な[Hawk2の要件](#)に加え、次の前提条件も満たす必要があります。

前提条件

- Hawk5の[ダッシュボード]で監視するすべてのクラスタでは、SUSE Linux Enterprise High Availability Extension 12 SP2を実行している必要があります。
- すべてのクラスタノードにあるHawk2の自己署名証明書を独自の証明書(または公式認証局によって署名された証明書)で置き換えていない場合は、「すべての」クラスタの「すべての」ノードで、少なくとも1回はHawk2にログインします。証明書を検証します(または、ブラウザで例外を追加して警告をスキップします)。そうしない場合、Hawk2はクラスタに接続できません。

手順 7.28: ダッシュボードを使用した複数クラスタの監視

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーで、[Dashboard (ダッシュボード)]を選択します。

Hawk2は、現在のクラスタサイトのリソースおよびノードに関する概要を表示します。また、Geoクラスタでの使用向けに設定された[チケット]を表示します。このビューで使用されているアイコンについての情報が必要な場合は、[凡例]をクリックします。リソースIDを検索するには、[検索]テキストボックスに名前(ID)を入力します。特定のノードのみを表示するには、フィルタアイコンをクリックしてフィルタリングオプションを選択します。

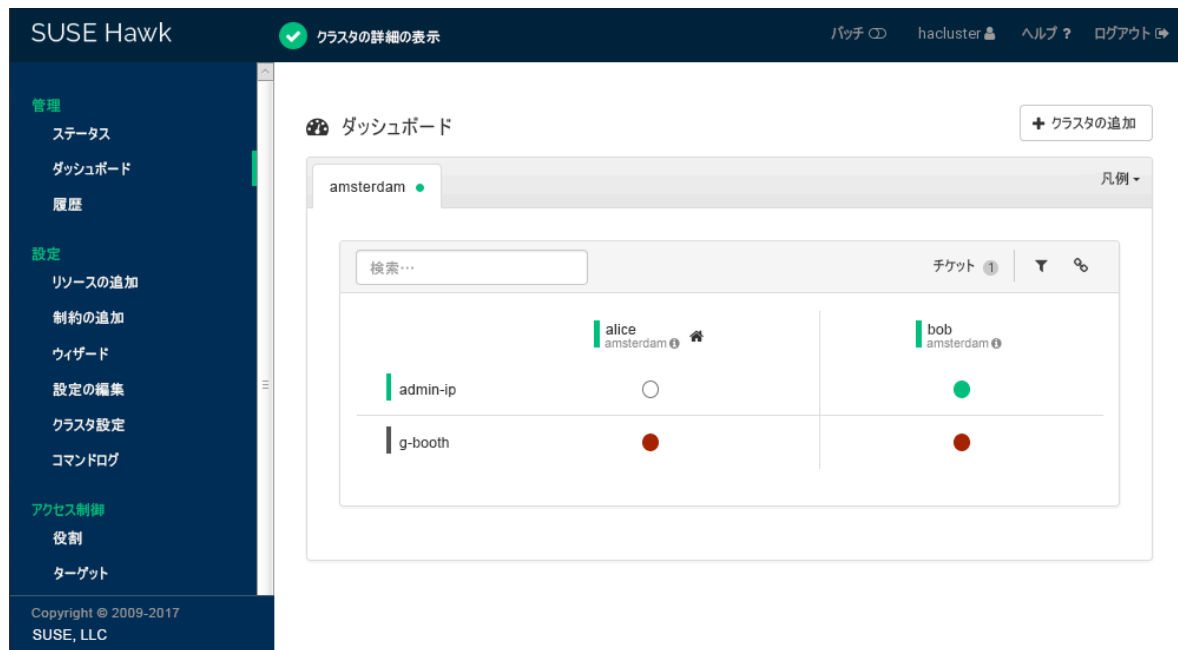


図 7.16: 1クラスタサイトのHAWK2ダッシュボード(amsterdam)

3. 複数のクラスタにダッシュボードを追加するには、以下の操作を行います。
 - a. [クラスタの追加]をクリックします。
 - b. [ダッシュボード]でクラスタを識別するためのクラスタ名を[クラスタ名]に入力します。たとえば、berlinです。
 - c. いずれかのノードの完全修飾ホスト名を、2つ目のクラスタに入力します。たとえば、charlieです。




- d. [追加]をクリックします。新たに追加されたクラスタサイトに対して、Hawk2は2つ目のタブにノードやリソースの概要を表示します。

注記: 接続エラー

パスワードを入力してノードにログインすることを求めるプロンプトが表示された場合、このノードにはまだ接続したことがなく、自己署名証明書を置き換えていない状態です。そのような場合は、パスワードを入力しても、接続は次のメッセージを表示して失敗します。

```
Error connecting to server. Retrying every 5 seconds...
```

続行するには、[自己署名証明書の置き換え](#)を参照してください。

4. クラスタサイトやその管理に関する詳細を参照するには、サイトのタブに切り替えてチェーンアイコンをクリックします。
Hawk2はこのサイトの[ステータス]ビューを新しいブラウザウィンドウかタブに表示します。この部分のGeoクラスタをそこから管理できます。
5. ダッシュボードからクラスタを削除するには、クラスタの詳細の右側にある  アイコンをクリックします。

7.9 バッチモードの使用

Hawk2は、「クラスタシミュレータ」を含む[バッチモード]を提供します。これは次の操作に使用できません。

- 各変更を直ちに反映させるのではなく、クラスタに変更をステージングして、それらの変更を単一トランザクションとして適用する。
- たとえば、潜在的な障害シナリオを調べるため、変更やクラスタイベントをシミュレートする。

たとえば、互いに依存するリソースのグループを作成する場合にバッチモードを使用できます。バッチモードを使用すると、クラスタに中間的なまたは不完全な設定を適用することを回避できます。

バッチモードが有効な間は、リソースや制約を追加したり編集したり、クラスタ設定を変更できます。ノードのオンラインまたはオフライン化、リソース操作およびチケットの許可または取り消しなど、クラスタのイベントをシミュレートすることもできます。詳細については、[手順7.30「ノード、リソース、またはチケットイベントの挿入」](#)を参照してください。

「クラスタシミュレータ」は、すべての変更後に自動的に実行され、ユーザインタフェースに予想される結果を表示します。たとえば、次のような場合もあります。バッチモードの最中にリソースを止めると、実際リソースはまだ実行中であるにも関わらず、ユーザインタフェースにはリソースが停止したと表示されます。

！ 重要: ウィザード、およびライブシステムへの変更

一部のウィザードには単なるクラスタ設定を超えるアクションが含まれます。これらのウィザードをバッチモードで使用する場合、クラスタ設定を超えるすべての変更がライブシステムに直ちに適用されます。

したがって、root 許可が必要なウィザードはバッチモードでは実行できません。

手順 7.29: バッチモードの使用

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. バッチモードを有効にするには、最上位の行から[バッチ]を選択します。
最上位の行の下に追加のバーが表示されます。これは、バッチモードがアクティブであることを示し、バッチモードで実行可能なアクションへのリンクが含まれます。



図 7.17: HAWK2バッチモードが有効

3. バッチモードがアクティブなときに、リソースや制約の追加または編集、あるいはクラスタ設定の編集など、クラスタに対する変更を実行します。
変更はシミュレートされ、すべての画面に表示されます。
4. 行った変更の詳細を表示するには、バッチモードバーから[表示]を選択します。[バッチモード]ウィンドウが開きます。
設定の変更について、ライブ状態とシミュレートされた変更間の相違がcrmsd構文で表示されます。
- 文字で開始される行は、現在の状態を表し、+で開始される行は、提案される状態を示します。
5. イベントを注入したり、さらに詳細を表示する場合は、[手順 7.30](#)を参照してください。または、[Close (閉じる)]をクリックしてウィンドウを閉じます。
6. シミュレートした変更を[破棄]または[適用]するかのいずれかを選択し、選択内容を確認します。これによりバッチモードが無効になり、通常のモードに戻ります。

バッチモードで実行中に、Hawk2では[Node Events (ノードイベント)]と[Resource Events (リソースイベント)]を注入することもできます。

[Node Events (ノードイベント)]

ノードの状態を変更できます。使用可能な状態は、[online (オンライン)]、[offline (オフライン)]、および[unclean (アンクリーン)]です。

[Resource Events (リソースイベント)]

リソースの一部のプロパティを変更できます。たとえば、操作(start、stop、monitorなど)、その操作を適用するノード、およびシミュレートされる予想される結果を設定できます。

[チケットイベント]

(Geoクラスタにおける)チケットの許可と取り消しの影響をテストできます。

手順 7.30: ノード、リソース、またはチケットイベントの挿入

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. バッチモードがまだアクティブでない場合、最上位の行で[バッチ]をクリックして、バッチモードに切り替えます。
3. バッチモードバーで、[表示]をクリックして、[バッチモード]ウィンドウを開きます。
4. ノードのステータスの変更をシミュレートするには
 - a. [注入] > [ノードイベント]を順にクリックします。
 - b. 操作する[ノード]を選択し、そのターゲット[状態]を選択します。
 - c. 変更内容を確認します。イベントは[バッチモード]ダイアログに一覧表示されるイベントのキューに追加されます。
5. リソースの操作をシミュレートするには
 - a. [注入] > [リソースイベント]を順にクリックします。
 - b. 操作する[リソース]を選択し、シミュレートする[操作]を選択します。
 - c. 必要に応じて、[間隔]を定義します。
 - d. 操作を実行する[ノード]を選択し、ターゲットとする[結果]を選択します。イベントは[バッチモード]ダイアログに一覧表示されるイベントのキューに追加されます。
 - e. 変更内容を確認します。
6. チケットアクションをシミュレートするには
 - a. [挿入] > [チケットイベント]を順にクリックします。
 - b. 操作する[チケット]を選択し、シミュレートする[アクション]を選択します。
 - c. 変更内容を確認します。イベントは[バッチモード]ダイアログに一覧表示されるイベントのキューに追加されます。
7. [バッチモード]ダイアログ(図 7.18)で、注入されたイベントごとに新しい行が表示されます。ここに一覧表示されるイベントは、直ちにシミュレートされ、[状態]画面に反映されます。

設定の変更も行った場合は、ライブ状態とシミュレートされた変更の間の相違が注入されたイベントの下に表示されます。



図 7.18: HAWK2のバッチモードで注入されたイベントと設定の変更

8. 注入されたイベントを削除するには、その隣の[削除]アイコンをクリックします。Hawk2では、それによって[状態]画面が更新されます。
9. 実行されたシミュレーションに関する詳細を表示するには、[シミュレータ]をクリックして、次のいずれかを選択します。

[概要]

詳細な概要を表示します。

[CIB (in)]/[CIB (out)]

[CIB (in)]では、初期のCIB状態を示します。[CIB (out)]では、遷移後のCIBの状態を示します。

[Transition Graph (遷移グラフ)]

遷移のグラフィカルな表現を示します。

[遷移]

遷移のXML表示を示します。

10. シミュレートされた変更を確認したら、[バッチモード]ウィンドウを閉じます。
11. バッチモードを終了するには、シミュレートされた変更を[適用]するか、[破棄]するか of the どちらかを選択します。

7.10 クラスタ履歴の表示

Hawk2には、クラスタの過去のイベントを表示する、次のような機能があります(いくつかの詳細さのレベルがあります)。

- 7.10.1項「ノードまたリソースの最近のイベントの表示」
- 7.10.2項「クラスタレポートのための履歴エクスプローラーの使用」
- 7.10.3項「履歴エクスプローラーの遷移詳細の表示」

7.10.1 ノードまたリソースの最近のイベントの表示

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーで、[Status (状態)]を選択します。[Resources (リソース)]と[Nodes (ノード)]が一覧表示されます。
3. リソースの最近のイベントを表示するには
 - a. [Resources (リソース)]をクリックして、それぞれのリソースを選択します。
 - b. リソースの[操作]列で、下矢印ボタンをクリックして[最近のイベント]を選択します。
Hawk2では、新しいウィンドウが開き、最新のイベントのテーブルビューが表示されます。
4. ノードの最近のイベントを表示するには
 - a. [Nodes (ノード)]をクリックして、それぞれのノードを選択します。
 - b. ノードの[操作]列で、[最近のイベント]を選択します。
Hawk2では、新しいウィンドウが開き、最新のイベントのテーブルビューが表示されます。

🔄 最近のイベント: alice

×

RC	リソース	操作	前回変更日	状態	コール	実行	完了
<u>0</u>	dummy1	dummy1_start_0	2016年10月25日火曜日18:10:49	開始しました	18	20ms	✓
<u>0</u>	dummy1	dummy1_monitor_10000	2016年10月25日火曜日18:10:49	開始しました	19	26ms	✓
<u>0</u>	dummy2	dummy2_stop_0	2016年10月25日火曜日18:10:49	停止 (無効)	15	23ms	✓
<u>0</u>	dummy2	dummy2_monitor_10000	2016年10月25日火曜日18:10:49	停止 (無効)	13	19ms	✓

7.10.2 クラスタレポートのための履歴エクスプローラーの使用

左のナビゲーションバーから[履歴]を選択して、[履歴エクスプローラー]にアクセスします。[履歴エクスプローラー]では、詳細なクラスタレポートを作成し、遷移情報を表示できます。次のオプションが表示されます。

[生成]

特定の時刻のクラスタレポートを作成します。Hawk2では、`crm report` コマンドをコールして、レポートを生成します。

[アップロード]

crmシェルスで直接作成したか、異なるクラスタ上で作成した `crm report` アーカイブをアップロードできます。

レポートを生成するか、アップロードした後で、これらのレポートは[レポート]の下に表示されます。レポートのリストから、レポートの詳細を表示したり、レポートをダウンロードしたり、削除したりできます。



図 7.19: HAWK2 - 履歴エクスプローラーのメインビュー

手順 7.31: クラスタレポートの生成またはアップロード

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーで、[History (履歴)]を選択します。
[生成]ビューに[履歴エクスプローラー]画面が開きます。デフォルトでは、レポートの提案される時間フレームは最後の時刻です。

3. クラスタレポートを作成するには

- レポートを直ちに開始するには、[生成]をクリックします。
- レポートの時間フレームを変更するには、提案される時間フレームの任意の場所をクリックし、ドロップダウンボックスから別のオプションを選択します。[Custom (カスタム)] 開始日、終了日、および時間をそれぞれ入力することもできます。レポートを開始するには、[生成]をクリックします。
レポートを終了した後で、そのレポートが[レポート]の下に表示されます。

4. クラスタレポートをアップロードするには、Hawk2でアクセス可能なファイルシステム上に `crm report` アーカイブがある必要があります。次の手順に従います。

- [アップロード] タブに切り替えます。
- クラスタレポートアーカイブを[ブラウズ]し、[アップロード]をクリックします。
レポートがアップロードされると、そのレポートが[レポート]の下に表示されます。

5. レポートをダウンロードまたは削除するには、[操作] 列のレポートの横の各アイコンをクリックします。

6. 履歴エクスプローラーのレポート詳細を表示するには、レポートの名前をクリックするか、[操作] 列から[表示]を選択します。

The screenshot shows the SUSE Hawk web interface. The sidebar on the left contains navigation links: 管理 (Management), ステータス (Status), ダッシュボード (Dashboard), 履歴 (History), 設定 (Settings), リソースの追加 (Add Resources), 制約の追加 (Add Constraints), ウィザード (Wizard), 設定の編集 (Edit Settings), クラスタ設定 (Cluster Settings), コマンドログ (Command Log), アクセス制御 (Access Control), 役割 (Roles), and ターゲット (Targets). The main content area is titled '履歴エクスプローラー' (History Explorer). It displays details for a report named 'hawk_2017_08_14T09_21_34_00_00_2017_08_14T15_21_34_00_00'. The report details include: 名前 (Name), 開始 (Start: 2017-08-14 09:23:19 UTC), 終了 (End: 2017-08-14 15:21:31 UTC), and 遷移 (Transition: 18). Below the details is a timeline visualization showing resource usage across nodes. The 'ノードイベント' (Node Events) and 'リソースイベント' (Resource Events) sections are also visible, displaying logs of system actions and resource state changes.

7. [レポート] ボタンをクリックして、レポートのリストに戻ります。

履歴エクスプローラーのレポート詳細

- レポートの名前です。
- レポートの開始時刻。
- レポートの終了時刻。
- レポートによってカバーされるクラスタのすべての遷移の遷移数およびタイムライン。遷移の詳細を表示する方法については、7.10.3項を参照してください。
- ノードイベント。
- リソースイベント。

7.10.3 履歴エクスプローラーの遷移詳細の表示

各遷移ごとに、クラスタはポリシーエンジン(PE)への入力として提供される状態のコピーを保存します。このアーカイブがログ記録されるパス。すべての pe-input* ファイルが指定コーディネータ(DC)上に生成されます。DCはクラスタ内で変更可能なため、いくつかのノードからの pe-input* ファイルがある場合があります。すべての pe-input* ファイルにPEの実行「予定」の内容が表示されます。

Hawk2では、各 pe-input* ファイルの名前とそれが作成された時刻とノードを表示できます。また、[履歴エクスプローラー]では、それぞれの pe-input* ファイルに基づいて、次の詳細をビジュアル化できます。

履歴エクスプローラーでの遷移詳細

[詳細]

遷移に属するログインデータのスニペットを表示します。次のコマンドの出力を表示します(リソースエージェントのログメッセージを含む)。

```
crm history transition peinput
```

[環境設定]

pe-input* ファイルが作成された時刻でクラスタ設定を表示します。

[差分]

選択した pe-input* ファイルと次のファイル間の設定とステータスの相違を表示します。

[ログ]

遷移に属するログインデータのスニペットを表示します。次のコマンドの出力を表示します。

```
crm history transition log peinput
```

これには、pengine、crmd、および lrmd からの詳細が含まれます。

[グラフ]

遷移のグラフィカルな表現を示します。[グラフ]をクリックすると、(pe-input* ファイルを使用して) PEが再度呼び出され、遷移のグラフィカルな表現が生成されます。

手順 7.32: 遷移詳細の表示

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーで、[History (履歴)]を選択します。
レポートがすでに生成またはアップロードされている場合は、これらのレポートが[Reports (レポート)]のリストに表示されます。そうでない場合は、[手順 7.31](#)で説明されるように、レポートを生成またはアップロードします。
3. レポート名をクリックするか、[操作]列から[表示]を選択して、[履歴エクスプローラーのレポート詳細](#)を開きます。
4. 遷移詳細にアクセスするには、下に示す遷移タイムラインの遷移ポイントを選択する必要があります。[Previous (前へ)]および[Next (次へ)]アイコンをおよび[Zoom In (拡大)]および[Zoom Out (縮小)]アイコンを使用して、興味ある遷移を探します。
5. pe-input* ファイルの名前、それが作成された時刻およびノードを表示するには、タイムラインの遷移ポイント上でマウスポインタを合わせます。
6. [履歴エクスプローラーでの遷移詳細](#)を表示するには、さらに知りたい遷移ポイントをクリックします。
7. [詳細]、[環境設定]、[差分]、[Logs (ログ)]または[Graph (グラフ)]を表示するには、[履歴エクスプローラーでの遷移詳細](#)で説明されるように、コンテンツを表示するには、各ボタンをクリックします。
8. レポートのリストに戻るには、[レポート]ボタンをクリックします。

7.11 クラスタヘルスの確認

Hawk2は、クラスタの検査と問題点の検出を行うウィザードを提供します。分析が完了すると、Hawk2は詳細なクラスタレポートを作成します。Hawk2によるクラスタヘルスの確認とレポートの生成には、ノード間のパスワード不要のSSHアクセスが必要です。ない場合は、現在のノードのデータのみを収集

します。[ha-cluster-bootstrap](#) パッケージが提供するブートストラップスクリプトでクラスタを設定した場合、パスワード不要のSSHアクセスは既に設定されています。手動で設定する必要がある場合は、[D.2項「パスワード不要のSSHアカウントの設定」](#)を参照してください。

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーから、[ウィザード]を選択します。
3. [基本的]カテゴリを展開します。
4. [Verify health and configuration (ヘルスと設定の検証)]ウィザードを選択します。
5. [検証]を選択して確認します。
6. クラスタのルートパスワードを入力して、[適用]をクリックします。Hawk2がレポートを生成します。

8 クラスタリソースの設定と管理(コマンドライン)

クラスタリソースを設定および管理するには、crmシェル(crmsh)コマンドラインユーティリティ、HA Web Konsole (Hawk2)、Webベースユーザインタフェースのいずれかを使用します。

この章では、`crm` コマンドラインツールを紹介し、このツールの概要、テンプレートの使用方法、そして、主にクラスタリソースの設定と管理(基本的なリソースと高度なリソース(グループとクローン)の作成、制約の設定、フェールオーバーノードとフェールバックノードの指定、リソース監視の設定、リソースの開始、クリーンアップ、または削除、および手動によるリソースの移行について説明します。



注記: ユーザの権限

クラスタを管理するには十分な権限が必要です。`crm` コマンドおよびそのサブコマンドは、`root` ユーザとして、またはCRM所有者ユーザとして実行される必要があります(通常は `hacluster` ユーザ)。

ただし、`user` オプションを使用することで、`crm` とそのサブコマンドを一般(権限のない)ユーザとして実行し、必要な場合はいつでも `sudo` を使用してIDを変更できます。たとえば、次の `crm` コマンドは、権限のあるユーザIDとして `hacluster` を使用します。

```
root # crm options user hacluster
```

`/etc/sudoers` を設定しておいて、`sudo` がパスワードを要求しないようにしておく必要があることに注意してください。

8.1 crmsh - 概要

`crm` コマンドには、リソース、CIB、ノード、リソースエージェントなどを管理するサブコマンドがあります。このコマンドには、例を組み込んだ詳細なヘルプシステムが用意されています。すべての例は、付録Bで説明される命名規則に従います。



ヒント: 対話型crmプロンプト

`crm` を引数なしで(または1つのサブレベルのみを引数として)使用することにより、crmシェルは対話モードになります。このモードは、次のプロンプトで示されます。

```
crm(live/HOSTNAME)
```

読みやすくするために、このマニュアルでは対話型crmのプロンプトでホスト名を省略します。次の例のように、aliceなどの特定のノードで対話型シェルを実行する必要がある場合にのみホスト名を含めます。

```
crm(live/alice)
```

8.1.1 ヘルプの表示

ヘルプには複数の方法でアクセスできます。

- `crm`とそのコマンドラインオプションの使用方法を出力するには:

```
root # crm --help
```

- 使用可能なすべてのコマンドの一覧を表示するには:

```
root # crm help
```

- コマンドの参照情報だけでなく、他のヘルプセクションにアクセスするには:

```
root # crm help topics
```

- `configure` サブコマンドの詳細なヘルプテキストを表示するには:

```
root # crm configure help
```

- `configure` の `group` サブコマンドの構文、使用方法、例を印刷するには:

```
root # crm configure help group
```

これも同様です:

```
root # crm help configure group
```

`help` サブコマンド(`--help` オプションと混同しないこと)のほとんどすべての出力によって、テキストビューアが開きます。このテキストビューアは上下にスクロール可能で、ヘルプテキストが読みやすくなっています。テキストビューアを閉じるには、`Q` キーを押します。



ヒント: バッシュおよび対話型シェルでタブ補完機能を使用

crmshは、対話型シェルに対してだけではなく、バッシュでの直接的で完全なタブ補完機能をサポートしています。たとえば、「`crm help config` `<Tab>`」と入力してを押すと、対話型シェルと同様に単語が補完されます。

8.1.2 crmshのサブコマンドの実行

`crm` コマンドそのものは、次のように使用できます。

- **直接:** すべてのサブコマンドを `crm` に続け、`Enter` を押すと、ただちにその出力が表示されます。たとえば、`crm help ra` を入力すると、`ra` サブコマンド(リソースエージェント)に関する情報を取得できます。
サブコマンドは、その短縮形が固有である限り短縮できます。たとえば、`status` を `st` と短縮しても、crmshにはユーザが意図したサブコマンドとして認識されます。
パラメータを短縮する機能もあります。通常、パラメータは `params` キーワードを使用して追加します。`params` セクションが最初のセクションでほかにセクションがない場合、このセクションを省略できます。たとえば、次のような行があるとして。

```
root # crm primitive ipaddr ocf:heartbeat:IPaddr2 params ip=192.168.0.55
```

これは次の行と同等です。

```
root # crm primitive ipaddr ocf:heartbeat:IPaddr2 ip=192.168.0.55
```

- **crmシェルスクリプトとして使用:** Crmシェルスクリプトには `crm` のサブコマンドが含まれます。詳細については、[8.1.4項「crmshのシェルスクリプトの使用」](#)を参照してください。
- **crmshクラスタスクリプトとして使用:** これらは、メタデータ、RPMパッケージへの参照、設定ファイル、およびcrmshサブコマンドを1つのわかりやすい名前バンドルしてまとめたものです。`crm script` コマンドを使用して管理します。
これらをcrmshシェルスクリプトと混同しないでください。両方に共通する目的はいくつかありますが、crmシェルスクリプトにはサブコマンドのみが含まれるのに対し、クラスタスクリプトにはコマンドの単純なエミュレーション以上の処理が組み込まれています。詳細については、[8.1.5項「crmshのクラスタスクリプトの使用」](#)を参照してください。
- **内部シェルとして対話式に使用:** 「`crm`」とタイプして、内部シェルに入ります。プロンプトが `crm(live)` に変化します。`help` を使用すると、利用可能なサブコマンドの概要を取得できます。内部シェルにはさまざまなサブコマンドレベルがあり、1つのサブコマンドをタイプして `Enter` を押すことで、そのサブコマンドのレベルに「入る」ことができます。

たとえば、「`resource`」とタイプすると、リソース管理レベルに入ります。プロンプトは `crm(live)resource#` に変わります。内部シェルを終了したい場合は、コマンド `quit`、`bye`、または `exit` を使用します。1レベル戻る場合は、`back`、`up`、`end`、または `cd` を使用します。`crm`、そしてオプションを付けずにサブコマンドを入力して `Enter` を押すと、そのレベルに直接入ることができます。

内部シェルは、サブコマンドとリソースのタブによる完了もサポートします。コマンドの冒頭をタイプして `<Tab>` を押すと、`crm` がそのオブジェクトを完了します。

すでに説明した方法に加えて、`crmsd` は、同期コマンド実行もサポートしています。これを有効にするには、`-w` オプションを使用します。`crm` を `-w` なしで起動した場合でも、後ほどユーザ初期設定の `wait` を `yes` に設定すれば (`options wait yes`)、有効にすることができます。このオプションが有効化される場合、`crm` は遷移が終了するまで待機します。処理が開始すると毎回、進行状況を示すための点が表示されます。同期コマンドの実行は `resource start` などのコマンドにのみ適用できます。



注記: 管理サブコマンドと設定サブコマンド間の相違

`crm` ツールには管理機能(サブコマンド `resources` および `node`)があり、設定に使用できます (`cib`、`configure`)。

以降のサブセクションでは、`crm` ツールの重要な側面について、その概要を示します。

8.1.3 OCFリソースエージェントに関する情報の表示

リソースエージェントはクラスタ設定で常に操作する必要があるため、`crm` ツールには、`ra` コマンドが含まれています。このコマンドを使用して、リソースエージェントの情報を表示し、リソースエージェントを管理します(詳細は6.3.2項「サポートされるリソースエージェントクラス」も参照)。

```
root # crm ra
crm(live)ra#
```

コマンド `classes` は、すべてのクラスとプロバイダを一覧表示します。

```
crm(live)ra# classes
lsb
ocf / heartbeat linbit lvm2 ocfs2 pacemaker
service
stonith
systemd
```

クラス(およびプロバイダ)に使用できるすべてのリソースエージェントの概要を取得するには、`list` コマンドを使用します。

```
crm(live)ra# list ocf
AoEtarget      AudibleAlarm    CTDB             ClusterMon
Delay          Dummy           EvmsSCC          Evmsd
Filesystem     HealthCPU       HealthSMART      ICP
IPaddr         IPaddr2         IPsrcaddr        IPv6addr
LVM            LinuxSCSI       MailTo           ManageRAID
ManageVE       Pure-FTPd       Raid1            Route
SAPDatabase    SAPInstance     SendArp          ServeRAID
...
```

リソースエージェントの概要は、`info` で表示できます。

```
crm(live)ra# info ocf:linbit:drbd
This resource agent manages a DRBD* resource
as a master/slave resource. DRBD is a shared-nothing replicated storage
device. (ocf:linbit:drbd)

Master/Slave OCF Resource Agent for DRBD

Parameters (* denotes required, [] the default):

drbd_resource* (string): drbd resource name
    The name of the drbd resource from the drbd.conf file.

drbdconf (string, [/etc/drbd.conf]): Path to drbd.conf
    Full path to the drbd.conf file.

Operations' defaults (advisory minimum):

    start          timeout=240
    promote        timeout=90
    demote         timeout=90
    notify         timeout=90
    stop           timeout=100
    monitor_Slave_0 interval=20 timeout=20 start-delay=1m
    monitor_Master_0 interval=10 timeout=20 start-delay=1m
```

ビューアは、「`q`」を押すと終了できます。



ヒント: `crm`の直接使用

前の例では、`crm` コマンドの内部シェルを使用しました。ただし、必ずしも、それを使用する必要はありません。該当するサブコマンドを `crm` に追加すれば、同じ結果が得られます。たとえば、すべてのOCFリソースエージェントを一覧するには、シェルに「`crm ra list ocf`」を入力すれば済みます。

8.1.4 crmshのシェルスクリプトの使用

crmshシェルスクリプトは、crmshサブコマンドをファイル内に列挙する便利な方法を提供します。これにより、特定の行をコメントしたり、これらのコメントを後で再生したりするのが簡単になります。crmshシェルスクリプトには「crmshサブコマンドのみ」を含めることができることに注意してください。他のコマンドは許可されていません。

crmshシェルスクリプトを使用するには、その前に特定のコマンドを使用してファイルを作成してください。たとえば、次のファイルにはクラスタのステータスが出力され、すべてのノードのリストが提供されます。

例 8.1: 単純なCRMSHシェルスクリプト

```
# A small example file with some crm subcommands
status
node list
```

ハッシュ記号(`#`)で始まる行はコメントなので、無視されます。行が長すぎる場合は、行末にバックslash(`\`)を挿入します。可読性を向上させるため、特定のサブコマンドに属する行をインデントすることをお勧めします。

このスクリプトを使用するには、次の方法のいずれかを使用します。

```
root # crm -f example.cli
root # crm < example.cli
```

8.1.5 crmshのクラスタスクリプトの使用

すべてのクラスタノードから情報を収集して変更をすべて展開することは、鍵となるクラスタ管理タスクです。複数のノードで同じ手順を手動で実行するのはミスを起こしがちであるため、代わりにcrmshクラスタスクリプトを使用できます。

これらを「crmshシェルスクリプト」と混同しないでください(8.1.4項「crmshのシェルスクリプトの使用」で説明)。

crmshシェルスクリプトとは対照的に、クラスタスクリプトでは次のような追加のタスクを実行します。

- 特定のタスクに必要なソフトウェアをインストールする。
- 設定ファイルを作成または変更する。
- 情報を収集し、クラスタの潜在的な問題をレポートする。
- 変更をすべてのノードに展開する。

crmshクラスタスクリプトは、他のクラスタ管理ツールを置き換えるものではなく、クラスタ全体に対して統合化された方法でこれらのタスクを実行できるようにします。詳細については、<http://crmsh.github.io/scripts/>を参照してください。

8.1.5.1 使用法

利用可能なすべてのクラスタのリストを取得するには、次のコマンドを実行します。

```
root # crm script list
```

スクリプトのコンポーネントを表示するには、`show` コマンドと、クラスタスクリプトの名前を使用します。次に例を示します。

```
root # crm script show mailto
mailto (Basic)
MailTo

  This is a resource agent for MailTo. It sends email to a sysadmin
  whenever a takeover occurs.

1. Notifies recipients by email in the event of resource takeover

  id (required) (unique)
    Identifier for the cluster resource
  email (required)
    Email address
  subject
    Subject
```

`show` の出力には、タイトル、短い説明、および手順が含まれます。各手順は一連のステップに分かれており、これらのステップを指定された順序で実行します。

各ステップには、必須パラメータとオプションパラメータのリスト、および短い説明とそのデフォルト値が含まれます。

各クラスタスクリプトは、一連の共通パラメータを認識します。これらのパラメータは任意のスクリプトに渡すことができます。

表 8.1: 共通パラメータ

パラメータ	引数	説明
<u>action</u>	<u>INDEX</u>	設定した場合、1つのアクションのみを実行します(verifyによって返されたインデックス)。

パラメータ	引数	説明
<u>dry_run</u>	<u>BOOL</u>	設定した場合、実行のシミュレートのみを行います(デフォルト: no)。
<u>nodes</u>	<u>LIST</u>	スクリプト実行対象のノードのリスト。
<u>port</u>	<u>NUMBER</u>	接続先のポート。
<u>statefile</u>	<u>FILE</u>	シングルステップ実行の場合に、指定したファイルに状態を保存します。
<u>sudo</u>	<u>BOOL</u>	設定した場合、sudoパスワードを入力するようcrmによってプロンプトが表示され、必要に応じてsudoが使用されます(デフォルト: no)。
<u>timeout</u>	<u>NUMBER</u>	秒単位での実行タイムアウト(デフォルト: 600)。
<u>user</u>	<u>USER</u>	指定したユーザとしてスクリプトを実行します。

8.1.5.2 クラスタスクリプトの検証と実行

問題を避けるため、クラスタスクリプトを実行する前に、実行するアクションを確認してパラメータを検証します。クラスタスクリプトは一連のアクションを実行でき、さまざまな理由で失敗する可能性があります。そのため、実行前にパラメータを検証すると、問題の回避に役立ちます。

たとえば、mailtoリソースエージェントでは、固有の識別子と電子メールアドレスが必要です。これらのパラメータを検証するには、以下を実行します。

```
root # crm script verify mailto id=sysadmin email=tux@example.org
1. Ensure mail package is installed

    mailx

2. Configure cluster resources

    primitive sysadmin ocf:heartbeat:MailTo
```

```
email="tux@example.org"
op start timeout="10"
op stop timeout="10"
op monitor interval="10" timeout="10"

clone c-sysadmin sysadmin
```

`verify` は各ステップを出力し、指定したパラメータでプレースホルダを置き換えます。問題が見つかった場合、`verify` によってレポートされます。問題がなければ、`verify` コマンドを `run` に置き換えます。

```
root # crm script run mailto id=sysadmin email=tux@example.org
INFO: MailTo
INFO: Nodes: alice, bob
OK: Ensure mail package is installed
OK: Configure cluster resources
```

`crm status` を使用して、リソースがクラスタに統合されているかどうかを確認します。

```
root # crm status
[...]
Clone Set: c-sysadmin [sysadmin]
Started: [ alice bob ]
```

8.1.6 設定テンプレートの使用



注記: 非推奨に関する注意

設定テンプレートの使用は非推奨で、今後削除される予定です。設定テンプレートはクラスタスクリプトに置き換えられます。[8.1.5項「crmshのクラスタスクリプトの使用」](#)を参照してください。

設定テンプレートは、crmsh用の既成のクラスタ設定です。リソーステンプレート([8.4.3項「リソーステンプレートの作成」](#)の説明を参照)と混同しないでください。これらはクラスタ用のテンプレートで、crmシェル用ではありません。

設定テンプレートは、最小限の操作で、特定ユーザのニーズに合わせて調整できます。テンプレートで設定を作成する際には、警告メッセージでヒントが与えられます。これは、後から編集することができ、さらにカスタマイズできます。

次の手順は、簡単ですが機能的なApache設定を作成する方法を示しています。

1. `root` としてログインし、`crm` 対話型シェルを開始します。

```
root # crm configure
```

2. 設定テンプレートから新しい設定を作成します。

- a. `template` サブコマンドに切り替えます。

```
crm(live)configure# template
```

- b. 使用可能な設定テンプレートを一覧します。

```
crm(live)configure template# list templates
gfs2-base  filesystem  virtual-ip  apache  clvm      ocfs2      gfs2
```

- c. 必要な設定テンプレートを決めます。Apache設定が必要なので、`apache` テンプレートを
選択し、`g-intranet`と名付けます。

```
crm(live)configure template# new g-intranet apache
INFO: pulling in template apache
INFO: pulling in template virtual-ip
```

3. パラメータを定義します。

- a. 作成した設定を一覧表示します。

```
crm(live)configure template# list
g-intranet
```

- b. 入力が必要とする最小限の変更項目を表示します。

```
crm(live)configure template# show
ERROR: 23: required parameter ip not set
ERROR: 61: required parameter id not set
ERROR: 65: required parameter configfile not set
```

- c. 好みのテキストエディタを起動し、[ステップ 3.b](#)でエラーとして表示されたすべての行に入
力します。

```
crm(live)configure template# edit
```

4. 設定を表示し、設定が有効かどうか確認します(太字のテキストは、[ステップ 3.c](#)で入力した設定
によって異なります)。

```
crm(live)configure template# show
primitive virtual-ip ocf:heartbeat:IPaddr \
  params ip="192.168.1.101"
primitive apache ocf:heartbeat:apache \
  params configfile="/etc/apache2/httpd.conf"
  monitor apache 120s:60s
group g-intranet \
```

```
apache virtual-ip
```

5. 設定を適用します。

```
crm(live)configure template# apply
crm(live)configure# cd ..
crm(live)configure# show
```

6. 変更内容をCIBに送信します。

```
crm(live)configure# commit
```

詳細がわかっている場合は、コマンドをさらに簡素化できます。次のコマンドをシェルで使用して、上記の手順を要約できます。

```
root # crm configure template \
new g-intranet apache params \
configfile="/etc/apache2/httpd.conf" ip="192.168.1.101"
```

内部 `crm` シェルに入っている場合は、次のコマンドを使用します。

```
crm(live)configure template# new intranet apache params \
configfile="/etc/apache2/httpd.conf" ip="192.168.1.101"
```

ただし、このコマンドは、設定テンプレートから設定を作成するだけです。設定をCIBに適用したり、コミットすることはありません。

8.1.7 シャドーイング設定のテスト

シャドーイング設定は、異なる設定シナリオのテストに使用されます。複数のシャドウ設定を作成した場合は、1つ1つテストして変更を加えた影響を確認できます。

通常の処理は次のようになります。

1. `root` としてログインし、`crm` 対話型シェルを開始します。

```
root # crm configure
```

2. 新しいシャドウ設定を作成します。

```
crm(live)configure# cib new myNewConfig
INFO: myNewConfig shadow CIB created
```

シャドウCIBの名前を省略する場合は、一時名の `@tmp@` が作成されます。

- 現在のライブ設定をシャドウ設定にコピーする場合は、次のコマンドを使用します。コピーしない場合は、このステップをスキップします。

```
crm(myNewConfig)# cib reset myNewConfig
```

このコマンドを使用すると、既存のリソースを後から編集する場合に、簡単に編集できます。

- 通常どおり変更を行います。シャドウ設定の作成後は、すべての変更がシャドウ設定に適用されます。すべての変更を保存するには、次のコマンドを使用します。

```
crm(myNewConfig)# commit
```

- ライブクラスタ設定が再び必要な場合は、次のコマンドでライブ設定に戻ります。

```
crm(myNewConfig)configure# cib use live
crm(live)#
```

8.1.8 設定の変更のデバッグ

設定の変更をクラスタにロードする前に、変更内容を `ptest` でレビューすることを推奨します。`ptest` コマンドを指定すると、変更のコミットによって生じるアクションのダイアグラムを表示できます。ダイアグラムを表示するには、`graphviz` パッケージが必要です。次の例は監視操作を追加するスクリプトです。

```
root # crm configure
crm(live)configure# show fence-bob
primitive fence-bob stonith:apcsmart \
    params hostlist="bob"
crm(live)configure# monitor fence-bob 120m:60s
crm(live)configure# show changed
primitive fence-bob stonith:apcsmart \
    params hostlist="bob" \
    op monitor interval="120m" timeout="60s"
crm(live)configure# ptest
crm(live)configure# commit
```

8.1.9 クラスタダイアグラム

クラスタダイアグラムを出力するには、コマンド `crm configure graph` を使用します。これにより現在の設定が現在のウィンドウに表示されるので、X11が必要になります。

SVG (Scalable Vector Graphics)を使用する場合は、次のコマンドを使用します。

```
root # crm configure graph dot config.svg svg
```

8.2 Corosync設定の管理

Corosyncは、ほとんどのHAクラスタの下層にあるメッセージング層です。`corosync` サブコマンドは、Corosync設定を編集および管理するためのコマンドを提供します。

たとえば、クラスタのステータスを一覧表示するには、`status`を使用します。

```
root # crm corosync status
Printing ring status.
Local node ID 175704363
RING ID 0
      id      = 10.121.9.43
      status   = ring 0 active with no faults
Quorum information
-----
Date:          Thu May  8 16:41:56 2014
Quorum provider: corosync_votequorum
Nodes:         2
Node ID:       175704363
Ring ID:       4032
Quorate:       Yes

Votequorum information
-----
Expected votes: 2
Highest expected: 2
Total votes:    2
Quorum:         2
Flags:          Quorate

Membership information
-----
    Nodeid      Votes Name
175704363        1 alice.example.com (local)
175704619        1 bob.example.com
```

`diff` コマンドは非常に便利です。すべてのノード上のCorosync設定を比較し(別途記載のない場合)、それらの差異を出力します。

```
root # crm corosync diff
--- bob
+++ alice
@@ -46,2 +46,2 @@
-     expected_votes: 2
-     two_node: 1
+     expected_votes: 1
+     two_node: 0
```

詳細については、http://crmsh.nongnu.org/crm.8.html#cmdhelp_corosync を参照してください。

8.3 グローバルクラスタオプションの設定

グローバルクラスタオプションは、一定の状況下でのクラスタの動作を制御します。事前に定義されている値は、通常は、そのまま保持できます。ただし、クラスタの主要機能を正しく機能させるには、クラスタの基本的なセットアップ後に、次のパラメータを調整する必要があります。

手順 8.1: `crm`でグローバルクラスタオプションを変更する

1. `root`としてログインし、`crm`ツールを開始します。

```
root # crm configure
```

2. 次のコマンドを使用して、2ノードクラスタだけのオプションを設定します。

```
crm(live)configure# property no-quorum-policy=stop
crm(live)configure# property stonith-enabled=true
```



重要: STONITHがない場合はサポートなし

STONITHがないクラスタはサポートされません。

3. 変更内容を表示します。

```
crm(live)configure# show
property $id="cib-bootstrap-options" \
  dc-version="1.1.1-530add2a3721a0ecccb24660a97dbfdaa3e68f51" \
  cluster-infrastructure="corosync" \
  expected-quorum-votes="2" \
  no-quorum-policy="stop" \
  stonith-enabled="true"
```

4. 変更内容をコミットして終了します。

```
crm(live)configure# commit
crm(live)configure# exit
```

8.4 クラスタリソースの設定

クラスタの管理者は、クラスタ内のサーバ上の各リソースや、サーバ上で実行する各アプリケーションに対してクラスタリソースを作成する必要があります。クラスタリソースには、Webサイト、電子メールサーバ、データベース、ファイルシステム、仮想マシン、およびユーザが常時使用できるようにする他のサーバベースのアプリケーションまたはサービスなどが含まれます。

作成できるリソースタイプの概要については、[6.3.3項「リソースのタイプ」](#)を参照してください。

8.4.1 ファイルからのクラスタリソースのロード

設定の一部またはすべてをローカルファイルまたはネットワークURLからロードできます。次の3つの異なる方法を定義できます。

置き換え

このオプションは、現在の設定を新たなソース設定に置き換えます。

update

このオプションは、ソース設定のインポートを試みます。現在の設定に新たな項目を追加したり、既存の項目を更新したりします。

push

このオプションは、ソースからのコンテンツを現在の設定にインポートします(updateと同じ)。ただし、新しい設定で使用できないオブジェクトを削除します。

ファイル mycluster-config.txt から新しい設定をロードするには、次の構文を使用します。

```
root # crm configure load push mycluster-config.txt
```

8.4.2 クラスタリソースの作成

クラスタで利用できるRA(リソースエージェント)には3種類あります(背景情報については6.3.2項「[サポートされるリソースエージェントクラス](#)」を参照)。新しいリソースをクラスタに追加するには、次の手順に従います。

1. root としてログインし、crm ツールを開始します。

```
root # crm configure
```

2. プリミティブIPアドレスを設定します。

```
crm(live)configure# primitive myIP ocf:heartbeat:IPaddr \  
    params ip=127.0.0.99 op monitor interval=60s
```

前のコマンドは「プリミティブ」に名前 myIP を設定します。クラス(ここでは ocf)、プロバイダ(heartbeat)、およびタイプ(IPaddr)を選択する必要があります。さらに、このプリミティブでは、IPアドレスなどのパラメータが必要です。自分の設定に合わせてアドレスを変更してください。

3. 行った変更を表示して確認します。

```
crm(live)configure# show
```

4. 変更をコミットして反映させます。

```
crm(live)configure# commit
```

8.4.3 リソーステンプレートの作成

類似した設定で複数のリソースを作成する場合、リソーステンプレートを使用すれば作業が簡単になります。基本的なバックグラウンド情報については、6.5.3項「リソーステンプレートと制約」を参照してください。これらを、「通常の」テンプレート(8.1.6項「設定テンプレートの使用」で説明したもの)と混同しないでください。次の構文を知るには、`rsc_template` コマンドを使用してください。

```
root # crm configure rsc_template
usage: rsc_template <name> [<class>:[<provider>:]]<type>
      [params <param>=<value> [<param>=<value>...]]
      [meta <attribute>=<value> [<attribute>=<value>...]]
      [utilization <attribute>=<value> [<attribute>=<value>...]]
      [operations id_spec
        [op op_type [<attribute>=<value>...] ...]]
```

たとえば、次のコマンドは、`ocf:heartbeat:Xen` リソースと、デフォルト値および操作に由来する `BigVM` の名前を持つ新しいリソーステンプレートを作成します。

```
crm(live)configure# rsc_template BigVM ocf:heartbeat:Xen \
  params allow_mem_management="true" \
  op monitor timeout=60s interval=15s \
  op stop timeout=10m \
  op start timeout=10m
```

新しいリソーステンプレートを定義したら、それをプリミティブとして使用すること、または順序、コロケーション、または `rsc_ticket` の制約で参照することができます。リソーステンプレートを参照するには、`@` 記号を使用します。

```
crm(live)configure# primitive MyVM1 @BigVM \
  params xmf="/etc/xen/shared-vm/MyVM1" name="MyVM1"
```

新しいプリミティブ `My-VM1` は、`BigVM` リソーステンプレートからすべてを継承します。たとえば、上の2つに等しいものは次のようになります。

```
crm(live)configure# primitive MyVM1 ocf:heartbeat:Xen \
  params xmf="/etc/xen/shared-vm/MyVM1" name="MyVM1" \
  params allow_mem_management="true" \
  op monitor timeout=60s interval=15s \
  op stop timeout=10m \
  op start timeout=10m
```

オプションや操作を上書きしたい場合は、自分の(プリミティブの)定義を追加します。たとえば、次の新しいプリミティブMyVM2は監視操作のタイムアウトを2倍にしますが、その他はそのままに残します。

```
crm(live)configure# primitive MyVM2 @BigVM \  
  params xfile="/etc/xen/shared-vm/MyVM2" name="MyVM2" \  
  op monitor timeout=120s interval=30s
```

リソーステンプレートは、そのテンプレートから派生するすべてのプリミティブを表すものとして、制約で参照することができます。これにより、クラスタ設定をいっそう簡潔かつクリアに行うことができます。リソーステンプレートは、場所の制約を除くすべての制約から参照することができます。コロケーション制約には、複数のテンプレート参照を含めることはできません。

8.4.4 STONITHリソースの作成

`crm`からは、STONITHデバイスは単なる1つのリソースと認識されます。STONITHリソースを作成するには、次の手順に従います。

1. `root`としてログインし、`crm`対話型シェルを開始します。

```
root # crm configure
```

2. 次のコマンドで、すべてのSTONITHタイプのリストを取得します。

```
crm(live)# ra list stonith  
apcmaster          apcmastersnmp          apcsmart  
baytech            bladehpi                cyclades  
drac3              external/drac5          external/dracmc-telnet  
external/hetzner   external/hmchttp        external/ibmrta  
external/ibmrta-telnet external/ipmi            external/ippower9258  
external/kdumpcheck external/libvirt          external/nut  
external/rackpdu   external/riloe           external/sbd  
external/vcenter  external/vmware          external/xen0  
external/xen0-ha   fence_legacy             ibmhmc  
ipmilan            meatware                  nw_rpc100s  
rcd_serial         rps10                     suicide  
wti_mpc            wti_nps
```

3. 上記のリストからSTONITHタイプを選択し、利用できるオプションのリストを表示します。次のコマンドを実行します。

```
crm(live)# ra info stonith:external/ipmi  
IPMI STONITH external device (stonith:external/ipmi)  
  
ipmitool based power management. Apparently, the power off  
method of ipmitool is intercepted by ACPI which then makes  
a regular shutdown. If case of a split brain on a two-node
```

```
it may happen that no node survives. For two-node clusters
use only the reset method.
```

Parameters (* denotes required, [] the default):

```
hostname (string): Hostname
    The name of the host to be managed by this STONITH device.
...
```

4. `stonith` クラス、[ステップ 3](#)で選択したタイプ、および必要に応じて該当するパラメータを使用して、STONITHリソースを作成します。たとえば、次のコマンドを使用します。

```
crm(live)# configure
crm(live)configure# primitive my-stonith stonith:external/ipmi \
    params hostname="alice" \
    ipaddr="192.168.1.221" \
    userid="admin" passwd="secret" \
    op monitor interval=60m timeout=120s
```

8.4.5 リソース制約の設定

すべてのリソースを設定することは、ジョブのほんの一部です。クラスタが必要なすべてのリソースを認識しても、正しく処理できるとは限りません。たとえば、DRBDのスレーブノードにファイルシステムをマウントしないようにしてください(実際、DRBDでは失敗します)。このような情報をクラスタが利用できるように、制約を定義します。

制約の詳細については、[6.5項「リソースの制約」](#)を参照してください。

8.4.5.1 場所の制約

`location` コマンドは、リソースを実行できるノード、できないノード、または実行に適したノードを定義するものです。

この種類の制約は、各リソースに複数追加できます。すべての `location` 制約は、所定のリソースに関して評価されます。`fs1` というIDを持つリソースを `alice` という名前のノード上で実行するプリファレンスを100にする簡単な例を次に示します。

```
crm(live)configure# location loc-fs1 fs1 100: alice
```

もう1つの例は、`pingd`による場所の設定です。

```
crm(live)configure# primitive pingd pingd \
    params name=pingd dampen=5s multiplier=100 host_list="r1 r2"
crm(live)configure# location loc-node_pref internal_www \
    rule 50: #uname eq alice \
```

```
rule pingd: defined pingd
```

場所の制約のもう1つの使用例は、「リソースセット」としてのプリミティブのグループ化です。これは、たとえば、いくつかのリソースがネットワーク接続のping属性によって異なるときに役立つ場合があります。以前は、`-inf/ping` ルールを設定で何度も重複して指定する必要があったため、設定内容が不必要に複雑でした。

次の例では、リソースセット `loc-alice` を作成し、仮想IPアドレス `vip1` および `vip2` を参照します。

```
crm(live)configure# primitive vip1 ocf:heartbeat:IPaddr2 params ip=192.168.1.5
crm(live)configure# primitive vip2 ocf:heartbeat:IPaddr2 params ip=192.168.1.6
crm(live)configure# location loc-alice { vip1 vip2 } inf: alice
```

ある場合には、`location` コマンドでリソースパターンを使用すると、より効率的で便利です。リソースパターンは、2つのスラッシュ間の正規表現です。たとえば、前に示した仮想IPアドレスは、次とすべて一致させることができます。

```
crm(live)configure# location loc-alice /vip.*/ inf: alice
```

8.4.5.2 コロケーション制約

`colocation` コマンドは、同じホストまたは別のホストで実行するべきリソースを定義するために使用します。

常に同じノードで実行する必要があるリソース、または同じノードで実行してはならないリソースを定義する場合には、それぞれ+infまたは-infのスコアを設定することだけが可能です。無限大以外のスコアの使用も可能です。その場合、コロケーションはadvisoryと呼ばれ、衝突が発生したときに他のリソースが停止しないようにするため、クラスタがそれらの制約に従わないこともあります。

たとえば、IDが `filesystem_resource` と `nfs_group` のリソースを常に同じホストで実行するには、次の制約を使用します。

```
crm(live)configure# colocation nfs_on_filesystem inf: nfs_group filesystem_resource
```

マスタスレーブ構成では、現在のノートがマスタかどうかと、リソースをローカルに実行しているかどうかを把握することが必要です。

8.4.5.3 依存性なしのリソースセットのコロケーション

同じノード上にリソースのグループを配置できると便利な場合がありますが(コロケーション制約を定義)、リソース間で困難な依存性を持つことはありません。

同じノード上にリソースを配置するが、これらの一方に障害が発生した場合のアクションがない場合は、`weak-bond` コマンドを使用します。


```
root # crm configure assist weak-bond RES1 RES2
```

weak-bondの実装により、指定されたリソースを持つダミーリソースとコロケーション制約が自動的に作成されます。

8.4.5.4 順序の制約

order コマンドは、アクションのシーケンスを定義します。

リソースのアクションや操作の順序を指定することが必要な場合があります。たとえば、デバイスがシステムで利用できるようになるまで、ファイルシステムはマウントできません。順序の制約を使用して、開始、停止、マスタへの昇格など、別のリソースが特殊な条件を満たす直前または直後に、サービスを開始または停止できます。

順序の制約を設定するには、次のようなコマンドを **crm** シェルで使します。

```
crm(live)configure# order nfs_after_filesystem mandatory: filesystem_resource nfs_group
```

8.4.5.5 サンプル設定のための制約

このセクションで使用される例は、制約を追加しないと機能しません。すべてのリソースは、必ず、マスタであるDRBDリソースと同じマシンで実行される必要があります。DRBDリソースは、他のリソースが開始する前にマスタにする必要があります。マスタでないときに、drbdデバイスをマウントしようとすると失敗します。次の制約を満たす必要があります。

- ファイルシステムは、常に、DRBDリソースのマスタと同じノード上に存在する必要があります。

```
crm(live)configure# colocation filesystem_on_master inf: \  
filesystem_resource drbd_resource:Master
```

- NFSサーバとIPアドレスは、ファイルシステムと同じノードに存在する必要があります。

```
crm(live)configure# colocation nfs_with_fs inf: \  
nfs_group filesystem_resource
```

- NFSサーバとIPアドレスは、ファイルシステムがマウントされた後に開始されます。

```
crm(live)configure# order nfs_second mandatory: \  
filesystem_resource:start nfs_group
```

- ファイルシステムは、drbdリソースがこのノードのマスタに昇格した後にマウントされる必要があります。

```
crm(live)configure# order drbd_first inf: \  
drbd_resource:promote filesystem_resource:start
```

8.4.6 リソースフェールオーバーノードの指定

リソースフェールオーバーを判定するには、メタ属性migration-thresholdを使用します。すべてのノードで失敗回数がmigration-thresholdを超えている場合には、リソースは停止したままになります。例:

```
crm(live)configure# location rsc1-alice rsc1 100: alice
```

通常、rsc1はaliceで実行されます。そこで失敗すると、migration-thresholdがチェックされ、失敗回数と比較されます。失敗回数がmigration-threshold以上の場合、次の候補のノードにマイグレートします。

開始が失敗すると、`start-failure-is-fatal` オプションによっては、失敗回数がinfに設定されます。stopの失敗により、フェンシングが発生します。STONITHが定義されていない場合には、リソースは移行しません。

概要については、6.5.4項「フェールオーバーノード」を参照してください。

8.4.7 リソースフェールバックノードの指定(リソースの固着性)

ノードがオンライン状態に戻り、クラスタ内にある場合は、リソースが元のノードにフェールバックすることがあります。リソースを実行していたノードにリソースをフェールバックさせたくない場合や、リソースのフェールバック先として別のノードを指定する場合は、リソースの固着性の値を変更します。リソースの固着性は、リソースの作成時でも、その後も指定できます。

概要については、6.5.5項「フェールバックノード」を参照してください。

8.4.8 負荷インパクトに基づくリソース配置の設定

一部のリソースは、メモリの最小量など、特定の容量要件を持っています。要件が満たされていない場合、リソースは全く開始しないか、またはパフォーマンスを下げた状態で実行されます。

これを考慮に入れて、High Availability Extensionでは、次のパラメータを指定できます。

1. 一定のノードが提供する容量
2. 一定のリソースが要求する容量
3. リソースの配置に関する全体的なストラテジ

パラメータと設定の詳細な背景情報については、6.5.6項「[負荷インパクトに基づくリソースの配置](#)」を参照してください。

リソースの要件とノードが提供する容量を設定するには、使用属性を使用します。使用属性に任意の名前を付け、設定に必要なだけ名前/値のペアを定義します。いくつかのエージェントは、たとえば `VirtualDomain` などの使用を更新します。

次の例では、クラスタのノードとリソースの基本設定がすでに完了していることを想定しています。さらに、特定のノードが提供する容量と特定のリソースが必要とする容量を設定します。

手順 8.2: `crm` で使用属性を追加または変更する

1. `root` としてログインし、`crm` 対話型シェルを開始します。

```
root # crm configure
```

2. ノードが提供する容量を指定するには、次のコマンドを使用し、プレースホルダ `NODE_1` をノードの名前に置き換えます。

```
crm(live)configure# node NODE_1 utilization memory=16384 cpu=8
```

これらの値によって、`NODE_1` は16GBのメモリと8つのCPUコアをリソースに提供すると想定されます。

3. リソースが要求する容量を指定するには、次のコマンドを使用します。

```
crm(live)configure# primitive xen1 ocf:heartbeat:Xen ... \
    utilization memory=4096 cpu=4
```

これによって、リソースは `NODE_1` からの4,096のメモリ単位と4つのCPU単位を使用します。

4. `property` コマンドを使用して、配置ストラテジを設定します。

```
crm(live)configure# property ...
```

次の値を使用できます。

default (デフォルト値)

使用値は考慮しません。リソースは、場所のスコアに従って割り当てられます。スコアが同じであれば、リソースはノード間で均等に分散されます。

utilization

リソースの要件を満たすだけの空き容量がノードにあるかどうか決定する際に、利用率を確認します。ただし、負荷分散は、まだ、ノードに割り当てられたリソースの数に基づいて行われます。

minimal

リソースの要件を満たすだけの空き容量がノードにあるかどうか決定する際に、利用率を確認します。できるだけ少ない数のノードにリソースを集中しようとします(残りのノードの電力節約のため)。

balanced

リソースの要件を満たすだけの空き容量がノードにあるかどうか決定する際に、利用率を確認します。リソースを均等に分散して、リソースのパフォーマンスを最適化しようとします。



注記: リソース優先度の設定

使用できる配置ストラテジは、最善策であり、まだ、複雑なヒューリスティックソルバで、常に最適な割り当て結果を得るには至っていません。リソースの優先度を正しく設定して、最重要なリソースが最初にスケジュールされるようにしてください。

5. 変更をコミットしてから、crmshを終了します。

```
crm(live)configure# commit
```

次の例は、同等のノードから成る3ノードクラスと4つの仮想マシンを示しています。

```
crm(live)configure# node alice utilization memory="4000"
crm(live)configure# node bob utilization memory="4000"
crm(live)configure# node charlie utilization memory="4000"
crm(live)configure# primitive xenA ocf:heartbeat:Xen \
    utilization hv_memory="3500" meta priority="10" \
    params xmfile="/etc/xen/shared-vm/vm1"
crm(live)configure# primitive xenB ocf:heartbeat:Xen \
    utilization hv_memory="2000" meta priority="1" \
    params xmfile="/etc/xen/shared-vm/vm2"
crm(live)configure# primitive xenC ocf:heartbeat:Xen \
    utilization hv_memory="2000" meta priority="1" \
    params xmfile="/etc/xen/shared-vm/vm3"
crm(live)configure# primitive xenD ocf:heartbeat:Xen \
    utilization hv_memory="1000" meta priority="5" \
    params xmfile="/etc/xen/shared-vm/vm4"
crm(live)configure# property placement-strategy="minimal"
```

3ノードはすべてアクティブであり、まず、xenAがノードに配置され、次に、xenDが配置されます。xenBとxenCは、一緒に割り当てられるか、またはどちらか1つがxenDとともに割り当てられます。

1つのノードに障害が発生した場合、残りのノード上で利用できるメモリ合計が少なすぎて、これらのリソースすべてはホストできません。xenAは確実に割り当てられ、xenDも同様です。ただし、xenBとxenCは、そのどちらか1つしか割り当てられません。xenBとxenCの優先度は同等なので、結果はまだ未定義です。これを解決するためにも、どちらかに高い優先度を設定する必要があります。

8.4.9 リソース監視の設定

リソースを監視するには、2つの方法(`op` キーワードで監視処理を定義するか、`monitor` コマンドを使用するか)があります。次の例では、Apacheリソースを設定し、`op` キーワードの使用で 60秒ごとに監視します。

```
crm(live)configure# primitive apache apache \  
  params ... \  
  op monitor interval=60s timeout=30s
```

同じことを次のようにしても実行できます。

```
crm(live)configure# primitive apache apache \  
  params ...  
crm(live)configure# monitor apache 60s:30s
```

概要については、6.4項「リソース監視」を参照してください。

8.4.10 クラスタリソースグループの構成

クラスタの一般的な要素の1つは、一緒の場所で見つける必要のあるリソースのセットです。連続的に開始し、逆の順序で停止します。この設定を簡単にするため、グループのコンセプトをサポートしています。次の例では、2つのプリミティブ(IPアドレスと電子メールリソース)を作成します。

1. `crm` コマンドをシステム管理者として実行します。プロンプトが `crm(live)` に変化します。
2. プリミティブを設定します。

```
crm(live)# configure  
crm(live)configure# primitive Public-IP ocf:heartbeat:IPaddr \  
  params ip=1.2.3.4 id= Public-IP  
crm(live)configure# primitive Email systemd:postfix \  
  params id=Email
```

3. 該当するIDを使用して、正しい順序で、プリミティブをグループ化します。

```
crm(live)configure# group g-mailsvc Public-IP Email
```

グループメンバーの順序を変更するには、`configure` サブコマンドから `modgroup` コマンドを使用します。プリミティブの `Email` を `Public-IP` の前に移動するには、次のコマンドを使用します(このコマンドは機能のデモのみを目的としています)。

```
crm(live)configure# modgroup g-mailsvc add Email before Public-IP
```

グループ(`Email` など)からリソースを削除する場合には、このコマンドを使用します。

```
crm(live)configure# modgroup g-mailsvc remove Email
```

概要については、6.3.5.1項「グループ」を参照してください。

8.4.11 クローンリソースの設定

クローンは当初、IPアドレスのN個のインスタンスを開始し、負荷分散のためにクラスタ上に分散させる便利な方法と考えられていました。それらは、DLMとの統合、サブシステムおよびOCFS2のフェンシングなど、他の目的にも有効であることがわかってきました。どのようなリソースでも、リソースエージェントがサポートしていれば、クローン化できます。

クローンリソースの詳細については、6.3.5.2項「クローン」を参照してください。

8.4.11.1 匿名クローンリソースの作成

匿名クローンリソースを作成するには、まずプリミティブリソースを作成して、それを `clone` コマンドで指定することです。次の操作を実行してください：

1. `root` としてログインし、`crm` 対話型シェルを開始します。

```
root # crm configure
```

2. 次のように、プリミティブを設定します。

```
crm(live)configure# primitive Apache ocf:heartbeat:apache
```

3. プリミティブをクローンします。

```
crm(live)configure# clone cl-apache Apache
```

8.4.11.2 ステートフル/マルチステートクローンリソースの作成

マルチステートリソースは、クローンが得意とするところです。これにより、インスタンスを2つの動作モード(active/passive、primary/secondary、またはmaster/slave)のいずれかに設定できます。

ステートフルクローンリソースを作成するには、まずプリミティブリソースを作成してから、マルチステートリソースを作成します。マルチステートリソースは少なくとも、昇格および降格操作をサポートしている必要があります。

1. `root` としてログインし、`crm` 対話型シェルを開始します。

```
root # crm configure
```

2. プリミティブを作成します。必要に応じて間隔を変更します。

```
crm(live)configure# primitive my-rsc ocf:myCorp:myAppl \  
    op monitor interval=60 \  
    op monitor interval=61 role=Master
```

3. マルチステートリソースを作成します。

```
crm(live)configure# ms ms-rsc my-rsc
```

8.5 クラスタリソースの管理

`crm` ツールでは、クラスタリソースの設定が可能だけでなく、既存リソースを管理することもできます。移行のサブセクションで概要を示します。

8.5.1 クラスタリソースの表示

クラスタを管理するには、コマンド `crm configure show` で、クラスタ設定、グローバルオプション、プリミティブなどの現在のCIBオブジェクトを一覧表示します。

```
root # crm configure show  
node 178326192: alice  
node 178326448: bob  
primitive admin_addr IPAddr2 \  
    params ip=192.168.2.1 \  
    op monitor interval=10 timeout=20  
primitive stonith-sbd stonith:external/sbd \  
    params pcmk_delay_max=30  
property cib-bootstrap-options: \  
    have-watchdog=true \  
    dc-version=1.1.15-17.1-e174ec8 \  
    cluster-infrastructure=corosync \  
    cluster-name=hacluster \  
    stonith-enabled=true \  
    placement-strategy=balanced \  
    standby-mode=true  
rsc_defaults rsc-options: \  
    resource-stickiness=1 \  
    migration-threshold=3  
op_defaults op-options: \  
    timeout=600 \  

```



```
record-pending=true
```

多数のリソースがある場合、`show`の出力が冗長になります。出力を制限するには、リソースの名前を使用します。たとえば、プリミティブ `admin_addr` のみのプロパティを一覧表示するには、リソース名を `show` に付加します。

```
root # crm configure show admin_addr
primitive admin_addr IPAddr2 \
    params ip=192.168.2.1 \
    op monitor interval=10 timeout=20
```

ただし、特定のリソースの出力をさらに制限したい場合があります。これは、「フィルタ」を使用して実現できます。フィルタは特定のコンポーネントに出力を制限します。たとえば、ノードのみを一覧表示するには、`type:node` を使用します。

```
root # crm configure show type:node
node 178326192: alice
node 178326448: bob
```

プリミティブにも興味がある場合には、`or` オペレータを使用します。

```
root # crm configure show type:node or type:primitive
node 178326192: alice
node 178326448: bob
primitive admin_addr IPAddr2 \
    params ip=192.168.2.1 \
    op monitor interval=10 timeout=20
primitive stonith-sbd stonith:external/sbd \
    params pcmk_delay_max=30
```

さらに、特定の文字列で開始するオブジェクトを検索するには次の表記を使用します。

```
root # crm configure show type:primitive and and 'admin*'
primitive admin_addr IPAddr2 \
    params ip=192.168.2.1 \
    op monitor interval=10 timeout=20
```

使用可能なすべてのタイプを一覧表示するには、`crm configure show type:` と入力し、`<Tab>` キーを押します。Bash補完により、すべてのタイプのリストが表示されます。

8.5.2 新しいクラスタリソースの開始

新しいクラスタリソースを開始するには、そのIDが必要です。次の手順に従います。

1. `root` としてログインし、`crm` 対話型シェルを開始します。

```
root # crm
```


2. リソースレベルに切り替えます。

```
crm(live)# resource
```

3. `start` でリソースを開始し、`<Tab>` キーを押してすべての既知のリソースを表示します。

```
crm(live)resource# start ID
```

8.5.3 リソースのクリーンアップ

リソースは、失敗した場合は自動的に再起動しますが、失敗のたびにリソースの失敗回数が増加します。`migration-threshold` がそのリソースに設定されている場合は、失敗の数が移行しきい値に達すると、そのリソースはノードで実行できなくなります。

1. シェルを開いて、`root` ユーザとしてログインします。
2. すべてのリソースのリストを取得します。

```
root # crm resource list
...
Resource Group: dlm-clvm:1
    dlm:1 (ocf:pacemaker:controld) Started
    clvm:1 (ocf:heartbeat:clvm) Started
```

3. リソース `dlm` をクリーンアップするには、たとえば、以下の手順を実行します:

```
root # crm resource cleanup dlm
```

8.5.4 クラスタリソースの削除

次の手順に従って、クラスタリソースを削除します。

1. `root` としてログインし、`crm` 対話型シェルを開始します。

```
root # crm configure
```

2. 次のコマンドを実行して、リソースのリストを取得します。

```
crm(live)# resource status
```

たとえば、出力はこのようなになります(ここで、`myIP` はリソースの該当するID)。

```
myIP (ocf:IPaddr:heartbeat) ...
```

3. 該当するIDを持つリソースを削除します(これは、`commit`も含意します)。

```
crm(live)# configure delete YOUR_ID
```

4. 変更をコミットします。

```
crm(live)# configure commit
```

8.5.5 クラスタリソースのマイグレーション

リソースは、ハードウェアまたはソフトウェアに障害が発生した場合、クラスタ内の他のノードに自動的にフェールオーバー(つまり移行)するよう設定されていますが、Hawk2またはコマンドラインを使用して、手動でリソースを別のノードに移動することもできます。

この作業を行うには、`migrate` コマンドを使用します。たとえば、リソース `ipaddress1` を `bob` というクラスタノードに移行するには、次のコマンドを使用します。

```
root # crm resource
crm(live)resource# migrate ipaddress1 bob
```

8.5.6 リソースのグループ化/タグ付け

タグは、コロケーションの作成や関係の順序付けを行わずに、複数のリソースをただちに参照する方法です。これは、概念的に関連するリソースをグループ化するのに役立つ場合があります。たとえば、データベースに関連するいくつかのリソースがある場合、`databases` というタグを作成し、データベースに関連するすべてのリソースをこのタグに追加します。

```
root # crm configure tag databases: db1 db2 db3
```

これにより、1つのコマンドですべてを起動できます。

```
root # crm resource start databases
```

同様に、すべてを停止することもできます。

```
root # crm resource stop databases
```

8.5.7 ヘルスステータスの取得

クラスタまたはノードの「ヘルス」ステータスは、「スクリプト」というもので表示できます。スクリプトは、ヘルスだけに限らず各タスクを実行できます。ただし、このサブセクションでは、ヘルスステータスを取得する方法に焦点を当てます。

`health` コマンドに関するすべての詳細を取得するには、`describe` を使用します。

```
root # crm script describe health
```

このコマンドは、すべてのパラメータの説明とリスト、およびそのデフォルト値を示します。スクリプトを実行するには、`run` を使用します。

```
root # crm script run health
```

スイートから1つのステップのみを実行したい場合は、`describe` コマンドのStepsカテゴリで、使用可能なすべてのステップを一覧表示できます。

たとえば、次のコマンドは、`health` コマンドの最初のステップを実行します。さらなる調査のために、出力が `health.json` に保存されます。

```
root # crm script run health
statefile='health.json'
```

上記のコマンドは、`crm cluster health` でも実行できます。

スクリプトに関する追加情報を表示するには、<http://crmsh.github.io/scripts/>  を参照してください。

8.6 cib.xmlから独立したパスワードの設定

クラスタ設定にパスワードなどの機密の情報が含まれている場合、それらをローカルファイルに保存する必要があります。こうしておけば、これらのパラメータがログに記録されたり、サポートレポートに漏洩することはありません。

`secret` を使用する前に、リソースの概要を確認するため、`show` コマンドを実行しておくといでしょう。

```
root # crm configure show
primitive mydb ocf:heartbeat:mysql \
  params replication_user=admin ...
```

上記の `mydb` リソースに対してパスワードを設定するには、次のコマンドを使用します。

```
root # crm resource secret mydb set passwd linux
INFO: syncing /var/lib/heartbeat/lrm/secrets/mydb/passwd to [your node list]
```

次のように、保存されたパスワードが返されます。

```
root # crm resource secret mydb show passwd
linux
```

パラメータは、ノード間で同期する必要があることに注意してください。`crm resource secret` コマンドを使用すれば、この処理が実行されます。秘密のパラメータを管理する場合には、このコマンドを使用することを強く推奨します。

8.7 履歴情報の取得

クラスタの履歴の調査は複雑な作業です。この作業を簡素化するために、`crmsh`には `history` コマンドとそのサブコマンドが含まれています。これは、SSHが正しく設定されていることが前提となります。

それぞれのクラスタは、状態を移動し、リソースを移行し、または重要なプロセスを開始します。これらすべてのアクションは、`history` のサブコマンドによって取得できます。

デフォルトでは、すべての `history` コマンドは過去1時間のイベントを確認します。このタイムフレームを変更するには、`limit` サブコマンドを使用します。構文は次のとおりです。

```
root # crm history
crm(live)history# limit FROM_TIME [TO_TIME]
```

有効な例として、次のようなものが挙げられます。

```
limit 4:00pm ,
```

```
limit 16:00
```


どちらのコマンドも同じ意味で、今日の午後4時を表しています。

```
limit 2012/01/12 6pm
```

2012年1月12日の午後6時。

```
limit "Sun 5 20:46"
```

今年の今月の5日日曜日の午後8時46分。

その他の例とタイムフレームの作成方法については、<http://labix.org/python-dateutil>  を参照してください。

`info` サブコマンドでは、`crm report` によって使用されているすべてのパラメータが表示されます。

```
crm(live)history# info
Source: live
Period: 2012-01-12 14:10:56 - end
Nodes: alice
Groups:
Resources:
```

`crm report` を特定のパラメータに制限するには、サブコマンド `help` で使用可能なオプションを表示します。

詳細レベルに絞り込んでいくには、サブコマンド `detail` とレベル数を使用します。

```
crm(live)history# detail 1
```

数値が大きいほど、レポートが詳細になっていきます。デフォルト値は `0` (ゼロ) です。

ここまでのパラメータを設定したら、`log` を使用してログメッセージを表示します。

最後の遷移を表示するには、次のコマンドを使用します。

```
crm(live)history# transition -1
INFO: fetching new logs, please wait ...
```

このコマンドはログを取得し、`dotty` (`graphviz` パッケージから)を実行して、遷移グラフを表示します。シェルはログファイルを開きます。ログ内は、`↓` と `↑` カーソルキーでブラウズできます。

遷移グラフを表示する必要がない場合には、`nograph` オプションを使用します。

```
crm(live)history# transition -1 nograph
```

8.8 詳細

- `crm` マニュアルページ。
- アップストリームプロジェクトマニュアルにアクセスします(<http://crmsh.github.io/documentation>)。
- 詳しい例については、項目「Highly Available NFS Storage with DRBD and Pacemaker」を参照してください。

9 リソースエージェントの追加または変更

クラスタによる管理が必要なすべての作業は、リソースとして使用できなければなりません。主要なグループとして、リソースエージェントとSTONITHエージェントの2つがあります。両方のカテゴリで、エージェントの追加や所有が可能で、クラスタ機能を各自のニーズに合わせて拡張することができます。

9.1 STONITHエージェント

クラスタがノードの1つの誤動作を検出し、そのノードの削除が必要となることがあります。これをフェンシングと呼び、一般にSTONITHリソースで実行されます。



警告: 外部SSH/STONITHはサポートされていません

SSHが他のシステムの問題にどのように反応するかを知る方法はありません。このため、外部SSH/STONITHエージェント(`stonith:external/ssh`など)は、運用環境ではサポートされていません。テスト目的でこのようなエージェントをまだ使用する場合は、`libglue-devel` パッケージをインストールしてください。

現在使用可能なすべてのSTONITHデバイス(ソフトウェア側から)のリストを入手するには、`crm ra list stonith` コマンドを使用します。お気に入りのエージェントが見つからない場合は、`-devel` パッケージをインストールしてください。STONITHデバイスおよびリソースエージェントの詳細については、第10章「フェンシングとSTONITH」を参照してください。

今のところ、STONITHエージェントの作成に関するマニュアルはありません。新しいSTONITHエージェントを作成する場合は、`cluster-glue` パッケージのソースに提供されている例を参照してください。

9.2 OCFリソースエージェントの作成

`/usr/lib/ocf/resource.d/` で提供されているすべてのOCFリソースエージェントの詳細については、6.3.2項「サポートされるリソースエージェントクラス」を参照してください。各リソースエージェントは、それを制御する次の操作をサポートしている必要があります。

`start`

リソースを開始または有効化します。

stop

リソースを中止または無効化します。

status

リソースのステータスを返します。

monitor

`status`と同様ですが、予期しない状態もチェックします。

validate

リソースの設定を検証します。

meta-data

リソースエージェントの情報をXMLで返します。

OCF RAの作成方法の一般的な手順は、次のとおりです。

1. テンプレートとして、`/usr/lib/ocf/resource.d/pacemaker/Dummy` ファイルをロードします。
2. 新しいリソースエージェントごとに新しいサブディレクトリを作成して、名前が競合しないようにします。たとえばリソースグループ `kitchen` にリソース `coffee_machine` がある場合、このリソースを `/usr/lib/ocf/resource.d/kitchen/` ディレクトリに追加します。このRAにアクセスするには、コマンド `crm` を実行します。

```
root # crm configure primitive coffee_1 ocf:coffee_machine:kitchen ...
```

3. 異なるシェル関数を実装し、異なる名前ファイルを保存します。

OCFリソースエージェントの作成についての詳細は、<https://github.com/ClusterLabs/resource-agents/blob/master/doc/dev-guides/ra-dev-guide.asc> を参照してください。コンセプトの特別な情報については、第1章「製品の概要」を参照してください。

9.3 OCF戻りコードと障害回復

OCF仕様によると、アクションが返す必要がある出口コードの厳密な定義があります。クラスタは常に、予期される結果に対する戻りコードを確認します。結果が予期された値と一致しない場合、アクションは失敗したとみなされ、回復処理が開始されます。障害回復には3種類あります。

表 9.1: 障害回復の種類

回復の種類	説明	クラスタが行うアクション
soft	一時的なエラーが発生しました。	リソースを再起動するか、新しい場所に移動させます。
hard	一時的ではないエラーが発生しました。エラーは、現在のノードに固有の場合があります。	リソースを他の場所に移動して、現在のノードで再試行されないようにします。
fatal	すべてのクラスタノードに共通の、一時的ではないエラーが発生しました。これは、不正な設定が指定されたことを示しています。	リソースを停止して、どのクラスタノードでも開始されないようにします。

アクションが失敗したと想定して、次の表では、異なるOCF戻りコードを概説します。また、エラーコードを受け取った場合にクラスタが開始する回復の種類も示しています。

表 9.2: OCF戻りコード

OCF戻りコード	OCFエイリアス	説明	回復の種類
0	OCF_SUCCESS	成功。コマンドは正常に完了しました。これは、すべてのstart、stop、promote、demoteコマンドの予期された結果です。	soft
1	OCF_ERR_GENERIC	汎用の「問題が発生した」ことを示すエラーコード。	soft
2	OCF_ERR_ARGS	リソースの設定がこのマシンで有効ではありません(たとえば、ノード上に見つからない場所/ツールを参照している場合)。	hard
3	OCF_ERR_UNIMPLEMENTED	要求されたアクションは実行されていません。	hard
4	OCF_ERR_PERM	リソースエージェントに、作業を完了できるだけの権限がありません。	hard

OCF戻りコード	OCFエイリアス	説明	回復の種類
5	OCF_ERR_-INSTALLED	リソースが必要とするツールがこのコンピュータにインストールされていません。	hard
6	OCF_ERR_-CONFIGURED	リソースの設定が無効です(たとえば、必要なパラメータがないなど)。	fatal
7	OCF_NOT_-RUNNING	リソースが実行されていません。クラスタは、どのアクションについてもこれを返すリソースを停止しようとしません。 このOCF戻りコードはリソース回復を必要することも必要としないこともあります。予期されたりリソースの状態に依存します。予期されない場合は、 <u>soft</u> 回復を行います。	N/A
8	OCF_RUNNING_-MASTER	リソースはマスタモードで実行しています。	soft
9	OCF_FAILED_-MASTER	リソースはマスタモードですが、失敗しました。リソースは降格、停止され、再度開始されます(昇格されます)。	soft
その他	該当なし	カスタムエラーコード。	soft

10 フェンシングとSTONITH

フェンシングはHA(High Availability)向けコンピュータクラスタにおいて、非常に重要なコンセプトです。クラスタがノードの1つの誤動作を検出し、そのノードの削除が必要となることがあります。これをフェンシングと呼び、一般にSTONITHリソースで実行されます。フェンシングは、HAクラスタを既知の状態にするための方法として定義できます。

クラスタのすべてのリソースには、それぞれ状態が関連付けられています。たとえば、「リソースr1はaliceで起動されている」などです。HAクラスタでは、このような状態は「リソースr1はalice以外のすべてのノードで停止している」ことを示します。クラスタは各リソースが1つのノードでのみ起動されるようにするためです。各ノードはリソースに生じた変更を報告する必要があります。つまり、クラスタの状態は、リソースの状態とノードの状態の集まりです。

ノードまたはリソースの状態を十分に確定することができない場合には、フェンシングが発生します。クラスタが所定のノードで起こっていることを認識しない場合でも、フェンシングによって、そのノードが重要なリソースを実行しないようにできます。

10.1 フェンシングのクラス

フェンシングには、リソースレベルとノードレベルのフェンシングという、2つのクラスがあります。後者について、この章で主に説明します。

リソースレベルのフェンシング

リソースレベルのフェンシングにより、特定のリソースへの排他的アクセスが保証されます。この一般的な例として、SANファイバチャネルスイッチからのノードのゾーニングの変更(つまり、ノードのディスクへのアクセスのロックアウト)や、SCSI予約などの方法が挙げられます。例については、[11.10項「ストレージ保護のための追加メカニズム」](#)を参照してください。

ノードレベルのフェンシング

ノードレベルのフェンシングにより、障害が発生したノードから共有リソースに完全にアクセスできなくなります。このことは通常、そのノードをリセットする、または電源オフにするというような、極端な手段で行われます。

10.2 ノードレベルのフェンシング

Pacemakerクラスタにおけるノードレベルフェンシングの実装は、STONITH (Shoot The Other Node in the Head: 他のノードの即時強制終了)です。High Availability Extensionには `stonith` コマンドラインツールが付属し、これはクラスタ上のノードの電源をリモートでオフにする拡張インタフェースです。使用できるオプションの概要については、`stonith --help` を実行するか、または `stonith` のマニュアルページで詳細を参照してください。

10.2.1 STONITHデバイス

ノードレベルのフェンシングを使用するには、まず、フェンシングデバイスを用意する必要があります。High Availability ExtensionでサポートされているSTONITHデバイスのリストを取得するには、任意のノード上で次のコマンドのいずれかを実行します。

```
root # stonith -L
```

または

```
root # crm ra list stonith
```

STONITHデバイスは次のカテゴリに分類できます。

電源分配装置(PDU)

電源分配装置は、重要なネットワーク、サーバ、データセンター装置の電力と機能を管理する、重要な要素です。接続した装置のリモートロード監視と、個々のコンセントでリモート電源オン/オフのための電力制御を実行できます。

無停電電源装置(UPS)

電力会社からの電力の停電発生時に別個のソースから電力を供給することで、安定した電源から接続先の装置に緊急電力が提供されます。

ブレード電源制御デバイス

クラスタを一連のブレード上で実行している場合、ブレードエンクロージャの電源制御デバイスがフェンシングの唯一の候補となります。当然、このデバイスは1台のブレードコンピュータを管理できる必要があります。

ライトアウトデバイス

ライトアウトデバイス(IBM RSA、HP iLO、Dell DRAC)は急速に広まっており、既製コンピュータの標準装備になる可能性さえあります。ただし、ホスト(クラスタノード)と電源を共有する場合は、必要時にそれらが機能しない場合があります。ノードに電力が供給されないままでは、それを

制御するデバイスも役に立ちません。したがって、バッテリー駆動のライトアウトデバイスを使用することを強くお勧めします。これらのデバイスはネットワークでアクセスできるという別の側面があります。これはシングルポイント障害またはセキュリティの懸念事項を示唆している可能性があります。

テストデバイス

テストデバイスは、テスト専用に使われます。通常、ハードウェアにあまり負担をかけないようになっています。クラスタが運用に使われる前に、実際のフェンシングデバイスに交換される必要があります。

STONITHデバイスは、予算と使用するハードウェアの種類に応じて選択します。

10.2.2 STONITHの実装

SUSE® Linux Enterprise High Availability ExtensionでのSTONITHの実装は、2つのコンポーネントで構成されています。

stonithd

stonithdは、ローカルプロセスまたはネットワーク経由でアクセスできるデーモンです。stonithdは、フェンシング操作に対応するコマンド(rest, power-off, power-on)を受け入れます。フェンシングデバイスのステータスチェックも行います。

stonithdデーモンはCRM HAクラスタの各ノードで実行されます。DCノードで実行されるstonithdインスタンスは、CRMからフェンシング要求を受け取ります。目的のフェンシング操作を実行するのは、このインスタンスとその他のstonithdプログラムです。

STONITHプラグイン

サポートされているフェンシングデバイスごとに、そのデバイスを制御できるSTONITHプラグインがあります。STONITHプラグインはフェンシングデバイスへのインタフェースです。cluster-glue パッケージに付属するSTONITHプラグインは、各ノード上の /usr/lib64/stonith/plugins にあります(fence-agents パッケージもインストールしている場合、そのパッケージに付属する各種プラグインは、/usr/sbin/fence_* にインストールされています)。すべてのSTONITHプラグインはstonithdからは同一のものと認識されますが、フェンシングデバイスの性質を反映しているため、大きな違いがあります。

一部のプラグインは、複数のデバイスをサポートします。代表的な例は ipmilan (または external/ipmi) で、IPMIプロトコルを実装し、このプロトコルをサポートする任意のデバイスを制御できます。

10.3 STONITHのリソースと環境設定

フェンシングをセットアップするには、1つまたは複数のSTONITHリソースを設定する必要があります。stonithdデーモンでは設定は不要です。すべての設定はCIBに保存されます。STONITHリソースはクラス `stonith` のリソースです(6.3.2項「サポートされるリソースエージェントクラス」を参照)。STONITHリソースはSTONITHプラグインのCIBでの表現です。フェンシング操作の他、STONITHリソースはその他のリソースと同様、起動、停止、監視できます。STONITHリソースの開始または停止は、ノード上でSTONITHデバイスドライバのロードおよびアンロードが行われることを意味しています。開始と停止は管理上の操作であるため、フェンシングデバイス自体での操作にはなりません。ただし、監視は、デバイスのログイン操作になります(必要な場合にデバイスが動作していることを検証するため)。STONITHリソースが別のノードにフェールオーバーすると、対応するドライバがロードされて、現在のノードがSTONITHデバイスと通信できるようにされます。

STONITHリソースはその他のリソースと同様にして設定できます。これらの操作の詳細については、使用しているクラスタ管理ツールに応じて次のいずれかを参照してください。

- Hawk2: 7.5.6項「STONITHリソースの追加」
- crmsh: 8.4.4項「STONITHリソースの作成」

パラメータ(属性)のリストは、それぞれのSTONITHの種類に依存します。特定のデバイスのパラメータ一覧を表示するには、`stonith` コマンドを実行します。

```
stonith -t stonith-device-type -n
```

たとえば、`ibmhmcc` デバイスタイプのパラメータを表示するには、次のように入力します。

```
stonith -t ibmhmcc -n
```

デバイスの簡易ヘルプテキストを表示するには、`-h` オプションを使用します。

```
stonith -t stonith-device-type -h
```

10.3.1 STONITHリソースの設定例

以降では、`crm` コマンドラインツールの構文で作成された設定例を紹介します。これを適用するには、サンプルをテキストファイル(`sample.txt` など)に格納して、実行します。

```
root # crm < sample.txt
```

`crm` コマンドラインツールでのリソースの設定については、第8章「クラスタリソースの設定と管理(コマンドライン)」を参照してください。

例 10.1: IBM RSAライトアウトデバイスの設定

IBM RSAライトアウトデバイスは、次のようにして設定できます。

```

configure
primitive st-ibmrsa-1 stonith:external/ibmrsa-telnet \
params nodename=alice ip_address=192.168.0.101 \
username=USERNAME password=PASSWORD
primitive st-ibmrsa-2 stonith:external/ibmrsa-telnet \
params nodename=bob ip_address=192.168.0.102 \
username=USERNAME password=PASSWORD
location l-st-alice st-ibmrsa-1 -inf: alice
location l-st-bob st-ibmrsa-2 -inf: bob
commit

```

この例では、location制約が使用されていますが、それは、STONITH操作が常に一定の確率で失敗するためです。したがって、(実行側でもあるノード上の) STONITH操作は信頼できません。ノードがリセットされていない場合、フェンシング操作結果について通知を送信できません。これを実行する方法は、操作が成功すると仮定して事前に通知を送信するほかありません。ただし操作が失敗した場合、問題が発生することがあります。したがって、規則によってstonithdはホストの終了を拒否します。

例 10.2: UPSフェンシングデバイスの設定

UPSタイプのフェンシングデバイスの設定は、上記の例と似ています。詳細についてはここでは割愛します。すべてのUPSデバイスは、フェンシングのために、同じ機構を使用します。デバイスへのアクセス方法が異なる方法。古いUPSデバイスにはシリアルポートしかなく、通常、特別のシリアルケーブルを使用して1200ボーで接続していました。新型の多くは、まだシリアルポートがありますが、USBインタフェースまたはEthernetインタフェースも使用します。使用できる接続の種類は、プラグインが何をサポートしているかによります。

たとえば、`apcmaster` を `apcsmart` デバイスと、`stonith -t stonith-device-type -n` コマンドを使用して比較します。

```
stonith -t apcmaster -h
```

次の情報が返されます。

```

STONITH Device: apcmaster - APC MasterSwitch (via telnet)
NOTE: The APC MasterSwitch accepts only one (telnet)
connection/session a time. When one session is active,
subsequent attempts to connect to the MasterSwitch will fail.
For more information see http://www.apc.com/
List of valid parameter names for apcmaster STONITH device:
ipaddr
login
password

```

今度は次のコマンドを使用します。

```
stonith -t apcsmart -h
```

次の結果が得られます。


```
STONITH Device: apcsmart - APC Smart UPS
(via serial port - NOT USB!).
Works with higher-end APC UPSes, like
Back-UPS Pro, Smart-UPS, Matrix-UPS, etc.
(Smart-UPS may have to be >= Smart-UPS 700?).
See http://www.networkupstools.org/protocols/apcsmart.html
for protocol compatibility details.
For more information see http://www.apc.com/
List of valid parameter names for apcsmart STONITH device:
ttydev
hostlist
```

最初のプラグインは、ネットワークポートとtelnetプロトコルを持つAPC UPSをサポートします。2番目のプラグインはAPC SMARTプロトコルをシリアル回線で使用します。これは多数のAPC UPS製品ラインでサポートされているものです。

例 10.3: KDUMPデバイスの設定

Kdumpは特殊なフェンシングデバイスに属し、実際にはフェンシングデバイスとは正反対のものです。このプラグインは、ノードでカーネルダンプが進行中かどうかをチェックします。進行中であればtrueを返し、ノードがフェンシングされたかのように動作します。

Kdumpプラグインは、別の実際のSTONITHデバイスと共に使用する必要があります(たとえば、`external/ipmi` など)。フェンシングメカニズムが正常に機能するには、実際のSTONITHデバイスがトリガされる前にKdumpをチェックするよう指定する必要があります。次の手順で示すように、`crm configure fencing_topology`を使用して、フェンシングデバイスの順序を指定してください。

1. kdump機能を有効にしたノードをすべて監視するには、`stonith:fence_kdump` リソースエージェント(パッケージ `fence-agents` で提供)を使用します。構成の例については、以下のリソースを参照してください。

```
configure
primitive st-kdump stonith:fence_kdump \
  params nodename="alice "\ ❶
  pcmk_host_check="static-list" \
  pcmk_reboot_action="off" \
  pcmk_monitor_action="metadata" \
  pcmk_reboot_retries="1" \
  timeout="60"
commit
```

- ❶ 監視されるノードの名前。複数のノードを監視する必要がある場合は、追加のSTONITHリソースを設定します。特定のノードでフェンシングデバイスを使用しないようにするには、場所に対する制約を追加します。

フェンシングのアクションは、リソースのタイムアウトが経過すると始まります。

2. 各ノード上の `/etc/sysconfig/kdump` で、kdumpプロセスが完了したときにすべてのノードに通知が送信されるように `KDUMP_POSTSCRIPT` を設定します。次に例を示します。

```
/usr/lib/fence_kdump_send -i INTERVAL -p PORT -c 1 alice bob charlie [...]
```

kdumpが完了すると、kdumpを実行するノードが自動的に再起動します。

3. ネットワークが有効化された `fence_kdump_send` ライブラリに関する指定を含む、新しい `initrd` を記述します。 `-f` オプションを使用して既存のファイルを上書きし、次のブートプロセスでその新規ファイルが使用されるようにします。

```
root # dracut -f -a kdump
```

4. `fence_kdump` リソース用のポートをファイアウォールで開きます。デフォルトポートは `7410` です。
5. 実際のフェンシングメカニズム(`external/ipmi` など)をトリガする前にKdumpがチェックされるようにするため、次のような設定を使用します。

```
fencing_topology \  
  alice: kdump-node1 ipmi-node1 \  
  bob: kdump-node2 ipmi-node2
```

`fencing_topology` の詳細:

```
crm configure help fencing_topology
```

10.4 フェンシングデバイスの監視

他のリソースと同様に、STONITHクラスのエージェントは、ステータスのチェックのための監視操作もサポートします。



注記: STONITHリソースの監視

STONITHリソースの監視は定期的に行われますが、頻繁ではありません。ほとんどのデバイスでは、少なくとも1800秒(30分)の監視間隔があれば十分です。

フェンシングデバイスはHAクラスタの不可欠な要素ですが、使用する必要が少ないほど好都合です。ブロードキャストトラフィックが多すぎると、しばしば、電源管理装置が影響を受けます。1分間に10本程度の接続しか処理できないデバイスもあります。2つのクライアントが同時に接続しようすると、混乱するデバイスもあります。大部分は、同時に複数のセッションを処理できません。

したがって、通常、フェンシングデバイスのステータスは数時間ごとにチェックすれば十分です。フェンシング操作の実行が必要となり、電源スイッチが故障する可能性は小さいものです。

監視操作の設定方法の詳細については、8.4.9項「リソース監視の設定」を参照してください(コマンドラインアプローチについて説明されている)。

10.5 特殊なフェンシングデバイス

実際のSTONITHデバイスを操作するプラグインに加えて、特殊目的のSTONITHプラグインも存在します。



警告: テスト目的のみ

次に示すSTONITHプラグインの一部は、デモとテストだけを目的としています。次のデバイスは、現実のシナリオでは使用しないでください。使用すると、データが損なわれたり、予測できない結果が生じることがあります。

- external/ssh
- ssh

fence_kdump

このプラグインは、ノードでカーネルダンプが進行中かどうかをチェックします。進行中であれば true を返し、ノードがフェンシングされたかのように動作します。いずれにせよ、ダンプ中には、ノードはどのリソースも実行できません。これによって、すでにダウンしているがダンプ中(これは時間がかかります)であるノードのフェンシングを避けることができます。このプラグインは、別の実際のSTONITHデバイスとともに使用する必要があります。設定の詳細については、例10.3「kdumpデバイスの設定」を参照してください。

external/sbd

これは自己フェンシングデバイスです。共有ディスクに挿入されることがある、いわゆる「ポイズンピル」に反応します。共有ストレージ接続が失われた場合、このデバイスはノードの動作を停止します。このSTONITHエージェントを使用してストレージベースのフェンシングを実装する方法については、第11章、手順11.7「SBDを使用するようにクラスタを設定する」を参照してください。詳細については、<https://github.com/ClusterLabs/sbd> も参照してください。

！ 重要: external/sbdおよびDRBD

external/sbd フェンシングメカニズムは、SBDパーティションが各ノードから直接読み取れることを要求します。そのため、SBDパーティションではDRBD*デバイスを使用してはなりません。

ただし、SBDパーティションが階層配置または複製されていない共有ディスク上にある場合には、DRBDクラスタでフェンシングメカニズムを使用することはできます。

external/ssh

別のソフトウェアベースの「フェンシング」メカニズムです。ノードは、root として、パスワードなしでお互いにログインできる必要があります。このメカニズムは、1つのパラメータ hostlist をとり、ターゲットにするノードを指定します。これは、本当に障害のあるノードをリセットすることはできないので、実際のクラスタには使用しないでください。これは、テストとデモの目的にのみ使用します。これを共有ストレージに使用すると、データが破損します。

meatware

meatware ではユーザが操作を支援する必要があります。起動すると、meatware はノードのコンソールに表示されるCRIT重大度メッセージを記録します。その場合、オペレータはノードがダウンしていることを確認して、meatclient(8) コマンドを発行します。これにより meatware は、クラスタに対して、そのノードが機能しなくなっていることを伝えます。詳細については、/usr/share/doc/packages/cluster-glue/README.meatware を参照してください。

suicide

これはソフトウェアのみのデバイスで、reboot コマンドを使用して実行しているノードを再起動できます。これにはノードのオペレーティングシステムによる操作が必要で、特定の状況では失敗することがあります。したがって、このデバイスの使用は、極力避けてください。ただし、1ノードクラスタで使用する場合は安全です。

ディスクレスSBD

この設定は、共有ストレージなしのフェンシングメカニズムが必要なときに便利です。このディスクレスモードでは、SBDは共有デバイスに頼らず、ハードウェアウォッチドッグを使用してノードをフェンスします。ただし、ディスクレスSBDは2ノードクラスタ用のスプリットブレイシナリオには対応できません。そのため、ディスクレスSBDを使用するには、3以上のノードが必要です。

suicide は、「自分のホストを停止させない」というルールに対する唯一の例外です。

10.6 基本的な推奨事項

次の推奨事項のリストをチェックして、よく発生する間違いを避けてください。

- 複数の電源スイッチを並列に接続しないでください。
- STONITHデバイスとその設定をテストする際には、各ノードからプラグを1回抜いて、ノードのフェンシングが起らないことを検証してください。
- リソースのテストは負荷のかかった状態で行って、タイムアウト値が適切であるかどうかを検証してください。タイムアウト値が短すぎると、(不必要な)フェンシング操作がトリガされることがあります。詳細については、[6.3.9項「タイムアウト値」](#)を参照してください。
- セットアップでは適切なフェンシングデバイスを使用してください。詳細については、[10.5項「特殊なフェンシングデバイス」](#)も参照してください。
- 1つ以上のSTONITHリソースを設定します。デフォルトでは、グローバルクラスタオプション `stonith-enabled` は `true` に設定されています。STONITHリソースが定義されていない場合、クラスタはどのリソースの開始も拒否します。
- グローバルクラスタオプション `stonith-enabled` を `false` に設定しないでください。これは、次の理由によります。
 - STONITHが有効でないクラスタはサポートされていません。
 - DLM/OCFS2は、決して発生しないフェンシング操作を待機して、永久にブロックし続けます。
- グローバルクラスタオプション `startup-fencing` を `false` に設定しないでください。デフォルトでは、これは次の理由で `true` に設定されています。クラスタの起動時に、あるノードが不明な状態になっていると、そのノードは、ステータスが明らかにされるまでフェンシングされます。

10.7 詳細

[/usr/share/doc/packages/cluster-glue](#)

インストールしたシステムのこのディレクトリには、多数のSTONITHプラグインおよびデバイスのREADMEファイルが格納されています。

http://clusterlabs.org/pacemaker/doc/crm_fencing.html 

STONITHに関する情報です。

<http://www.clusterlabs.org/doc/> 

- 『Pacemakerの説明』: Pacemakerの設定に必要なコンセプトを説明します。包括的で詳しい参照情報です。

http://techthoughts.typepad.com/managing_computers/2007/10/split-brain-quo.html 

HAクラスタでのスプリットブレイン、クォーラム、フェンシングのコンセプトを説明する記事。

11 ストレージ保護とSBD

SBD (STONITH Block Device)は、共有ブロックストレージ(SAN、iSCSI、FCoEなど)を介したメッセージの交換を通じて、Pacemakerベースのクラスタのノードフェンシングメカニズムを提供します。これにより、フェンシングメカニズムが、ファームウェアバージョンの変更や特定のファームウェアコントローラへの依存から切り離されます。動作異常のノードが本当に停止したかどうかを確認するために、各ノードではウォッチドッグが必要です。特定の条件下では、ディスクレスモードで実行することにより、共有ストレージなしでSBDを使用することもできます。

`ha-cluster-bootstrap` スクリプトは、フェンシングメカニズムとしてSBDを使用するオプションを用いて、クラスタを設定する自動化された方法を提供します。詳細については、『インストールおよびセットアップクイックスタート』を参照してください。ただし、SBDを手動で設定する場合、個々の設定に関するオプションが増えます。

この章では、SBDの背後にある概念について説明します。スプリットブレインシナリオの場合に潜在的なデータ破損からクラスタを保護するために、SBDが必要とするコンポーネントを設定する手順を説明します。

ノードレベルのフェンシングに加えて、LVM2排他アクティブ化やOCFS2ファイルロックのサポート(リソースレベルのフェンシング)など、ストレージ保護のための追加のメカニズムを使用することができます。これにより、管理上またはアプリケーション上の障害からシステムが保護されます。

11.1 概念の概要

SBDは、「Storage-Based Death」または「STONITHブロックデバイス」の略語です。

High Availabilityクラスタスタックの最優先事項は、データの整合性を保護することです。これは、データストレージへの非協調的な同時アクセスを防止することによって実現されます。クラスタスタックは、複数の制御メカニズムを使用してこの処理を行います。

ただし、ネットワークのパーティション分割やソフトウェアの誤動作により、クラスタでいくつかのDCが選択される状況となる可能性があります。このいわゆるスプリットブレインシナリオが発生した場合は、データが破損することがあります。

STONITHによるノードフェンシングは、これを防ぐためのプライマリメカニズムです。ノードフェンシングメカニズムとしてSBDを使用することは、スプリットブレインシナリオの場合に、外部電源オフデバイスを使用せずにノードをシャットダウンする1つの方法です。

SBDパーティション

すべてのノードが共有ストレージへのアクセスを持つ環境で、デバイスの小さなパーティションをSBDでできるようにフォーマットします。パーティションのサイズは、使用されるディスクのブロックサイズによって異なります(たとえば、512バイトのブロックサイズの標準SCSIディスクには1MB、4KBブロックサイズのDASDディスクには4MB必要です)。初期化プロセスでは、最大255のノードに対するスロットを備えたデバイス上にメッセージレイアウトが作成されます。

SBDデーモン

SBDは、そのデーモンの設定後、クラスタスタックの他のコンポーネントが起動される前に各ノードでオンラインになります。SBDデーモンは、他のすべてのクラスタコンポーネントがシャットダウンされた後で終了されます。したがって、クラスタリソースがSBDの監督なしでアクティブになることはありません。

メッセージ

このデーモンは、自動的に、パーティション上のメッセージスロットの1つを自分自身に割り当て、自分へのメッセージがないかどうか、そのスロットを絶えず監視します。デーモンは、メッセージを受信すると、ただちに要求に従います(フェンシングのための電源切断や再起動サイクルの開始など)。

また、デーモンは、ストレージデバイスへの接続性を絶えず監視し、パーティションが到達不能になった場合は、デーモン自体が終了します。このため、デーモンがフェンシングメッセージから切断されることはありません。これは、クラスタデータが別のパーティション上の同じ論理ユニットにある場合、追加障害ポイントになることはありません。ストレージ接続を失えば、ワークロードは終了します。

ウォッチドッグ

SBDを使用する場合は常に、正常動作するウォッチドッグが不可欠です。近代的なシステムは、ソフトウェアコンポーネントによって「チックル」または「フィード」される必要のあるhardware watchdogをサポートします。ソフトウェアコンポーネント(この場合、SBDデーモン)は、ウォッチドッグにサービスパルスを定期的書き込むことによって、ウォッチドッグに「フィード」します。デーモンがウォッチドッグへのフィードを停止すると、ハードウェアでシステムが強制的に再起動されます。この機能は、SBDプロセス自体の障害(I/Oエラーで終了またはスタックするなど)に対する保護を提供します。

Pacemaker統合が有効になっている場合、デバイスの過半数が失われてもSBDはセルフフェンスを行います。たとえば、クラスタにA、B、Cの3つのノードが含まれており、ネットワーク分割によってAには自分自身しか表示できず、BとCはまだ通信可能な状態であるとして。この場合、2つのクラスタパーティションが存在し、1つは過半数(B、C)であるためにクォーラムがあり、もう1つにはクォーラムがない(A)ことになります。過半数のフェンシングデバイスに到達できないときにこれが発生した場合、ノードAはすぐに自らダウンしますが、BとCは引き続き実行されます。

11.2 SBDの手動設定の概要

手動でストレージベースのフェンシングを設定するには、次の手順に従う必要があります。これらは `root` として実行する必要があります。開始する前に、11.3項「要件」を確認してください。

1. ウォッチドッグのセットアップ
2. シナリオに応じて、1～3台のデバイスとともにまたはディスクレスモードでSBDを使用してください。概要については、11.4項「SBDデバイスの数」を参照してください。詳細な設定については、以下に記載されています。
 - デバイスでのSBDの設定
 - ディスクレスSBDの設定
3. SBDとフェンシングのテスト

11.3 要件

- ストレージベースのフェンシングには、最大3つのSBDデバイスを使用できます。1～3台のデバイスを使用する場合、共有ストレージにすべてのノードからアクセス可能である必要があります。
- 共有ストレージデバイスのパスが永続的で、クラスタ内のすべてのノードで一致している必要があります。`/dev/disk/by-id/dm-uuid-part1-mpath-abcdef12345` などの固定デバイス名を使用してください。
- 共有ストレージはFC (ファイバチャネル)、FCoE (Fibre Channel over Ethernet)、またはiSCSI経由で接続できます。仮想化環境では、ハイパーバイザーは共有ブロックデバイスを提供する場合があります。どの場合にも、共有ブロックデバイス上のコンテンツがすべてのクラスタノードに対して一貫性がある必要があります。キャッシュによってその一貫性が損なわれないようにしてください。
- 共有ストレージセグメントが、ホストベースのRAID、LVM2、またはDRBD*を「使用してはなりません」。DRBDは分割できますが、SBDでは2つの状態が存在することはできないため、これはSBDにとって問題になります。クラスタマルチデバイス(クラスタMD)は、SBDには使用できません。
- ただし、信頼性向上のため、ストレージベースのRAIDとマルチパスの使用は推奨されます。
- 255を超えるノードでデバイスを共有しない限り、異なるクラスタ間でSBDデバイスを共有できます。
- 3つ以上のノードがあるクラスタの場合は、SBDをディスクレスモードで使用することもできます。

11.4 SBDデバイスの数

SBDは、最大3つのデバイスの使用をサポートしています。

1台のデバイス

最も単純な実装です。すべてのデータが同じ共有ストレージ上にあるクラスタに適しています。

2台のデバイス

この設定は、主に、ホストベースのミラーリングを使用しているものの3つ目のストレージデバイスが使用できない環境で役立ちます。1つのミラーログにアクセスできなくなっても、SBDは終了せず、クラスタは引き続き実行できます。ただし、SBDにはストレージの非同期分割を検出できるだけの情報が与えられていないので、ミラーログが1つだけ使用可能なときにもう一方をフェンスすることはできません。つまり、ストレージアレイのいずれかがダウンしたときに、2つ目の障害に自動的に耐えることはできません。

3台のデバイス

最も信頼性の高い設定です。障害または保守による1台のデバイスの機能停止から回復できません。複数のデバイスが失われた場合、およびクラスタパーティションまたはノードの状態に応じて必要な場合にのみ、SBD自体が終了します。少なくとも2つのデバイスにまだアクセス可能な場合は、フェンシングメッセージを正常に送信できます。

この設定は、ストレージが1つのアレイに制約されていない、比較的複雑なシナリオに適しています。ホストベースのミラーリングソリューションでは、1つのミラーログに1つのSBDを設定し(自分自身はミラーしない)、iSCSI上に追加のタイブレーカを設定できます。

ディスクレス

この設定は、共有ストレージなしのフェンシングメカニズムが必要なときに便利です。このディスクレスモードでは、SBDは共有デバイスに頼らず、ハードウェアウォッチドッグを使用してノードをフェンスします。ただし、ディスクレスSBDは2ノードクラスタ用のスプリットブレインシナリオには対応できません。そのため、ディスクレスSBDを使用するには、3以上のノードが必要です。

11.5 タイムアウトの計算

フェンシングメカニズムとしてSBDを使用する場合、すべてのコンポーネントのタイムアウトを考慮することが重要です。それらのコンポーネントが相互に依存するためです。

ウォッチドッグのタイムアウト

このタイムアウトは、SBDデバイスの初期化中に設定されます。これは主にストレージのレイテンシに依存します。この時間内に大半のデバイスを正常に読み込む必要があります。それができない場合、そのノードでセルフフェンスを行うことがあります。



注記: マルチパスまたはiSCSIセットアップ

マルチパスセットアップまたはiSCSI上にSBDデバイスがある場合、パスの障害を検出して次のパスに切り替えるのに必要な時間に、タイムアウトを設定する必要があります。

またこれは、`/etc/multipath.conf` で `max_polling_interval` の値が ウォッチドッグ のタイムアウト未満でなければならないことを意味します。

msgwait タイムアウト

このタイムアウトは、SBDデバイスの初期化中に設定されます。この時間が経過するとSBDデバイス上のノードのスロットに書き込まれたメッセージが配信されたとみなされる時間を定義します。タイムアウトは、ノードでセルフフェンスを行う必要があることを検出するのに十分な長さでなければなりません。

ただし、msgwait タイムアウトが比較的長い場合、フェンシングアクションが戻る前にフェンスされたクラスタノードが再加入することがあります。これは、SBD設定の `SBD_DELAY_START` パラメータを設定することで軽減できます([手順 11.4のステップ 4](#)で説明)。

CIBの stonith-timeout

このタイムアウトは、グローバルクラスタプロパティとしてCIBで設定されます。これは、STONITHアクション(再起動、オン、オフ)が完了するのを待つ時間の長さを定義します。

CIBの stonith-watchdog-timeout

このタイムアウトは、グローバルクラスタプロパティとしてCIBで設定されます。明示的に設定されていない場合は、デフォルトで `0` に設定されます。これは1~3台のデバイスとともにSBDを使用するのに適しています。ディスクレスモードでSBDを使用する方法の詳細については、[手順 11.8「ディスクレスSBDの設定」](#)を参照してください。

ウォッチドッグのタイムアウトを変更する場合は、他の2つのタイムアウトも調整する必要があります。次の「式」は、これら3つの値の関係を示しています。

例 11.1: タイムアウト計算の式

```
Timeout (msgwait) >= (Timeout (watchdog) * 2)
stonith-timeout = Timeout (msgwait) + 20%
```

たとえば、ウォッチドッグのタイムアウトを `120` に設定した場合、msgwait タイムアウトを `240` に設定し、stonith-timeout を `288` に設定します。

`ha-cluster-bootstrap` スクリプトを使用してクラスタを設定し、SBDデバイスを初期化する場合、これらのタイムアウト間の関係が自動的に考慮されます。

11.6 ウォッチドッグのセットアップ

SUSE Linux Enterprise High Availability Extensionには、ハードウェア固有のウォッチドッグドライバを提供する、いくつかのカーネルモジュールが付属しています。最もよく使用されるカーネルモジュールのリストについては、よく使用されるウォッチドッグドライバを参照してください。

運用環境のクラスタでは、ハードウェア固有のウォッチドッグドライバを使用することをお勧めします。ただし、ハードウェアに適合するウォッチドッグがない場合、カーネルウォッチドッグモジュールとして `softdog` を使用することができます。

High Availability Extensionはウォッチドッグに「フィード」するソフトウェアコンポーネントとしてSBDデーモンを使用します。

11.6.1 ハードウェアウォッチドッグの使用

特定のシステムの正しいウォッチドッグカーネルモジュールを判断することは、容易ではありません。自動プロービングは頻繁に失敗します。その結果、正しいモジュールがロードされる前に、多くのモジュールがすでにロードされている状態になってしまいます。

表 11.1は、最もよく使用されるウォッチドッグドライバのリストです。お使いのハードウェアがそこに記載されていない場合、ディレクトリ `/lib/modules/KERNEL_VERSION/kernel/drivers/watchdog` も選択肢のリストとして用意されています。または、ハードウェアベンダーに名前を問い合わせてください。

表 11.1: よく使用されるウォッチドッグドライバ

Hardware (ハードウェア)	ドライバ
HP	<code>hpwdt</code>
Dell, Supermicro, Lenovo	<code>iTCO_wdt</code>
Fujitsu	<code>ipmi_watchdog</code>
IBMメインフレーム上のz/VMのVM	<code>vmwatchdog</code>
Xen VM (DomU)	<code>xen_wdt</code>
Generic	<code>softdog</code>

！ 重要: ウォッチドッグタイマへのアクセス

一部のハードウェアベンダーは、システムのリセット用にウォッチドッグを使用するシステム管理ソフトウェアを提供しています(たとえば、HP ASRデーモンなど)。ウォッチドッグがSBDで使用されている場合は、このようなソフトウェアを無効にします。他のソフトウェアは、ウォッチドッグタイマにアクセスしないでください。

手順 11.1: 正しいカーネルモジュールのロード

正しいウォッチドッグモジュールがロードされていることを確認するには、次の手順を実行します。

1. お使いのカーネルバージョンでインストールされているドライバをリストします。

```
root # rpm -ql kernel-VERSION | grep watchdog
```

2. カーネルに現在ロードされているウォッチドッグモジュールをリストします。

```
root # lsmod | egrep "(wd|dog)"
```

3. 結果が表示されたら、間違ったモジュールをアンロードします。

```
root # rmmod WRONG_MODULE
```

4. お使いのハードウェアに適合するウォッチドッグモジュールを有効にします。

```
root # echo WATCHDOG_MODULE > /etc/modules-load.d/watchdog.conf
root # systemctl restart systemd-modules-load
```

5. ウォッチドッグモジュールが正しくロードされているかどうかをテストします。

```
root # lsmod | egrep "(wd|dog)"
```

11.6.2 ソフトウェアウォッチドッグ(softdog)の使用

運用環境のクラスタでは、ハードウェア固有のウォッチドッグドライバを使用することをお勧めします。ただし、ハードウェアに適合するウォッチドッグがない場合、カーネルウォッチドッグモジュールとして softdog を使用することができます。

！ 重要: softdogの制限

softdogドライバはCPUが最低1つは動作中であることを前提とします。すべてのCPUが固まっている場合、システムを再起動させるsoftdogドライバのコードは実行されません。これに対して、ハードウェアウォッチドッグはすべてのCPUが固まっても動作し続けます。

手順 11.2: SOFTDOGカーネルモジュールのロード

1. softdogドライバを有効にします。

```
root # echo softdog > /etc/modules-load.d/watchdog.conf
```

2. `/etc/modules-load.d/watchdog.conf` に `softdog` モジュールを追加し、サービスを再起動します。

```
root # echo softdog > /etc/modules-load.d/watchdog.conf
root # systemctl restart systemd-modules-load
```

3. softdogウォッチドッグモジュールが正しくロードされているかどうかをテストします。

```
root # lsmod | grep softdog
```

11.7 デバイスでのSBDの設定

セットアップには次の手順が必要です。

1. SBDデバイスの初期化
2. SBD設定ファイルの編集
3. SBDサービスの有効化と起動
4. SBDデバイスのテスト
5. SBDを使用するようにクラスタを設定する

開始する前に、SBDに使用するブロックデバイスが、[11.3項](#)で指定された要件を満たしていることを確認してください。

SBDデバイスを設定するときは、いくつかのタイムアウト値を考慮する必要があります。詳細については、[11.5項「タイムアウトの計算」](#)を参照してください。

ノード上で実行しているSBDデーモンがウォッチドッグタイマを十分な速さで更新していない場合、ノード自体が終了します。タイムアウトを設定したら、個別の環境でテストしてください。

手順 11.3: SBDデバイスの初期化

共有ストレージでSBDを使用するには、まず1〜3台のブロックデバイス上でメッセージングレイアウトを作成する必要があります。`sbd create` コマンドは、指定された1つまたは複数のデバイスにメタデータヘッダを書き込みます。また、最大255ノードのメッセージングスロットを初期化します。追加のオプションを指定せずに実行する場合、このコマンドはデフォルトのタイムアウト設定を使用します。



警告: 既存データの上書き

SBD用に使用するデバイスには、重要なデータが一切ないようにしてください。`sbd create` コマンドを実行すると、指定されたブロックデバイスの最初のメガバイトが、さらなる要求やバックアップなしに上書きされます。

1. SBDに使用するブロックデバイスを決定します。
2. 次のコマンドで、SBDデバイスを初期化します。

```
root # sbd -d /dev/SBD create
```

(`/dev/SBD`を実際のパス名で置き換えます。たとえば `/dev/disk/by-id/scsi-ST2000DM001-0123456_Wabdefg` です)。

SBDに複数のデバイスを使用するには、`-d` オプションを複数回指定します。たとえば、次のようになります。

```
root # sbd -d /dev/SBD1 -d /dev/SBD2 -d /dev/SBD3 create
```

3. SBDデバイスがマルチパスグループにある場合は、`-1` と `-4` オプションを使用して、SBDに使用するタイムアウトを調整します。詳細については、[11.5項「タイムアウトの計算」](#)を参照してください。タイムアウトはすべて秒単位で指定します。

```
root # sbd -d /dev/SBD -4 180 ① -1 60 ② create
```

- ① `-4` オプションは `msgwait` タイムアウトを指定するために使用されます。上の例では、180 秒に設定されます。
- ② `-1` オプションは `watchdog` タイムアウトを指定するために使用されます。上の例では、60 秒に設定されます。エミュレートされたウォッチドッグで使用可能な最小値は 15 秒です。

4. デバイスに書き込まれた内容を確認します。

```
root # sbd -d /dev/SBD dump
Header version      : 2.1
UUID                : 619127f4-0e06-434c-84a0-ea82036e144c
Number of slots     : 255
Sector size        : 512
Timeout (watchdog)  : 60
Timeout (allocate) : 2
Timeout (loop)      : 1
Timeout (msgwait)   : 180
==Header on disk /dev/SBD is dumped
```

ご覧のように、タイムアウトがヘッダにも保存され、それらに関するすべての参加ノードの合意が確保されます。

SBDデバイスを初期化したら、SBD設定ファイルを編集し、次にそれぞれのサービスを有効にして起動し、変更を有効にします。

手順 11.4: SBD設定ファイルの編集

1. ファイル `/etc/sysconfig/sbd`を開きます。
2. 次のパラメータを検索します。 `SBD_DEVICE`
SBDメッセージを交換するために監視および使用するデバイスを指定します。
3. `SBD`をお使いのSBDデバイスに置き換えて、この行を編集します。

```
SBD_DEVICE="/dev/SBD"
```

1行目で複数のデバイスを指定する必要がある場合は、セミコロンで区切って指定します(デバイスの順序は任意で構いません)。

```
SBD_DEVICE="/dev/SBD1; /dev/SBD2; /dev/SBD3"
```

SBDデバイスがアクセス不能な場合は、SBDデーモンが開始できなくなり、クラスタの起動を抑制します。

4. 次のパラメータを検索します。 `SBD_DELAY_START`.
遅延を有効または無効にします。 `SBD_DELAY_START` を `yes` に設定します(`msgwait`が比較的長い場合)。ただし、クラスタノードは非常に高速に起動します。このパラメータを `yes` に設定すると、ブート時にSBDの起動が遅れます。これは、仮想マシンで必要となることがあります。

SBDデバイスをSBD設定ファイルに追加したら、SBDデーモンを有効にします。SBDデーモンは、クラスタスタックの不可欠なコンポーネントです。これは、クラスタスタックが実行されているときに、実行されている必要があります。したがって、`sbd` サービスは、`pacemaker` サービスが開始されるたびに依存関係として開始されます。

手順 11.5: SBDサービスの有効化と起動

1. 各ノードで、SBDサービスを有効にします。

```
root # systemctl enable sbd
```

これは、Pacemakerサービスが開始されるたびに、Corosyncサービスと一緒に開始されます。

2. 各ノードでクラスタスタックを再起動します。

```
root # systemctl stop pacemaker
root # systemctl start pacemaker
```

これによって、自動的にSBDデーモンの開始がトリガされます。

次の手順として、[手順 11.6](#)の説明に従ってSBDデバイスをテストします。

手順 11.6: SBDデバイスのテスト

1. 次のコマンドを使用すると、ノードスロットとそれらの現在のメッセージがSBDデバイスからダンプされます。

```
root # sbd -d /dev/SBD list
```

ここでSBDを使用して起動したすべてのクラスタノードが表示されます。たとえば、2ノードクラスタを使用している場合、両方のノードのメッセージスロットには clear と表示されます。

0	alice	clear
1	bob	clear

2. ノードの1つにテストメッセージを送信してみます。

```
root # sbd -d /dev/SBD message alice test
```

3. ノードがシステムログファイルにメッセージの受信を記録します。

```
May 03 16:08:31 alice sbd[66139]: /dev/SBD: notice: servant: Received command test
from bob on disk /dev/SBD
```

これによって、SBDがノード上で実際に機能し、メッセージを受信できることが確認されます。

最後のステップとして、[手順 11.7](#)の説明に従ってクラスタ設定を調整する必要があります。

手順 11.7: SBDを使用するようにクラスタを設定する

クラスタでSBDの使用を設定するには、クラスタ設定で次の操作を行う必要があります。

- 設定に適合する値に stonith-timeout パラメータを設定します。
- SBD STONITHリソースを設定します。

stonith-timeout の計算については、[11.5項「タイムアウトの計算」](#)を参照してください。

1. シェルを起動し、root または同等のものとしてログインします。
2. crm configure を実行します。
3. 次のように入力します。


```
crm(live)configure# property stonith-enabled="true" ❶  
crm(live)configure# property stonith-watchdog-timeout=0 ❷  
crm(live)configure# property stonith-timeout="220s" ❸
```

- ❶ STONITHを使用しないクラスタはサポートされていないため、これがデフォルト設定になります。ただし、テスト目的でSTONITHが無効化されている場合は、再度このパラメータが true に設定されていることを確認してください。
 - ❷ 明示的に設定されていない場合、この値はデフォルトで 0 に設定されます。これは1〜3台のデバイスとともにSBDを使用するのに適しています。
 - ❸ SBDの `msgwait` タイムアウト値が 30 秒に設定されていた場合、stonith-timeout 値は 220 が適切です。
4. 2ノードクラスタの場合、予測可能な遅延を希望するか、ランダムな遅延を希望するかを決めます。他のクラスタ設定については、このパラメータを設定する必要はありません。

予測可能な静的遅延

このパラメータはSTONITHアクションを実行する前に静的遅延を有効にします。別々のフェンシングリソースおよび異なる遅延値が使用されている場合に、ノードが互いにフェンシングしないようにします。対象ノードは「フェンシングの競合」で失われます。2ノードクラスタのスプリットブレインシナリオの場合に、このパラメータを使用して、特定のノードが存続するよう「マーク付けする」ことができます。これを正常に実行するには、各ノードに2つのプリミティブSTONITHデバイスを作成することが必須です。次の設定では、スプリットブレインシナリオの場合に、aliceが勝利して存続します。

```
crm(live)configure# primitive st-sbd-alice stonith:external/sbd params \  
    pcmk_host_list=alice pcmk_delay_base=20  
crm(live)configure# primitive st-sbd-bob stonith:external/sbd params \  
    pcmk_host_list=bob pcmk_delay_base=0
```

動的なランダム遅延

このパラメータは、SBDなどの低速デバイスを使用する場合の二重フェンシングを防止します。これは、フェンシングデバイスに対するSTONITHアクションのランダム遅延を追加します。スプリットブレインシナリオの場合、両方のノードが互いにフェンスを試みる可能性がある2ノードクラスタでは特に重要です。

```
crm(live)configure# primitive stonith_sbd stonith:external/sbd  
    params pcmk_delay_max=30
```

5. `show` で変更内容をレビューします。
6. `commit` で変更を送信し、`exit` でcrmライブ設定を終了します。

リソースの起動後、SBDを使用するためにクラスタが正常に設定されます。ノードをフェンスする必要がある場合にこの方法を使用します。

11.8 ディスクレスSBDの設定

ディスクレスモードでSBDを動作させることができます。このモードでは、次の場合にウォッチドッグデバイスを使用してノードをリセットします。クォーラムが失われた場合、監視されているデーモンが失われて回復しなかった場合、またはノードでフェンシングが必要であるとPacemakerが判断した場合。ディスクレスSBDは、クラスタの状態、クォーラム、およびいくつかの合理的な前提に応じた、ノードの「セルフフェンシング」に基づいています。STONITH SBDリソースプリミティブはCIBでは必要ありません。

！ 重要: クラスタノード数

2ノードクラスタのフェンシングメカニズムとしてディスクレスSBDを使用しないでください。3つ以上のノードを含むクラスタでのみ使用してください。ディスクレスモードのSBDでは、2ノードクラスタのスプリットブレインシナリオを処理できません。

手順 11.8: ディスクレスSBDの設定

1. ファイル `/etc/sysconfig/sbd` を開き、次のエントリを使用します。

```
SBD_PACEMAKER=yes
SBD_STARTMODE=always
SBD_DELAY_START=no
SBD_WATCHDOG_DEV=/dev/watchdog
SBD_WATCHDOG_TIMEOUT=5
```

共有ディスクが使用されていないので、`SBD_DEVICE` エントリは不要です。このパラメータがない場合、`sbd` サービスはSBDデバイスのウォッチャプロセスを開始しません。

2. 各ノードで、SBDサービスを有効にします。

```
root # systemctl enable sbd
```

これは、Pacemakerサービスが開始されるたびに、Corosyncサービスと一緒に開始されます。

3. 各ノードでクラスタスタックを再起動します。

```
root # systemctl stop pacemaker
root # systemctl start pacemaker
```

これによって、自動的にSBDデーモンの開始がトリガされます。

4. パラメータ `have-watchdog=true` が自動的に設定されているかどうかを確認します。

```
root # crm configure show | grep have-watchdog
have-watchdog=true
```

5. `crm configure` を実行し、crmシェルで次のクラスタプロパティを設定します。

```
crm(live)configure# property stonith-enabled="true" ❶
crm(live)configure# property stonith-watchdog-timeout=10 ❷
```

- ❶ STONITHを使用しないクラスタはサポートされていないため、これがデフォルト設定になります。ただし、テスト目的でSTONITHが無効化されている場合は、再度このパラメータが `true` に設定されていることを確認してください。
 - ❷ ディスクレスSBDの場合、このパラメータはゼロであってはなりません。これは、どれくらいの時間が経ったらフェンシングターゲットがすでにセルフフェンスを行ったとみなされるのかを定義します。したがって、その値は、`SBD_WATCHDOG_TIMEOUT` (`/etc/sysconfig/sbd` 内)の値以上である必要があります。SUSE Linux Enterprise High Availability Extension 15から、`stonith-watchdog-timeout` を負の値に設定した場合、Pacemakerは自動的にこのタイムアウトを計算し、`SBD_WATCHDOG_TIMEOUT` の値の2倍に設定します。
6. `show` で変更内容をレビューします。
7. `commit` で変更を送信し、`exit` でcrmライブ設定を終了します。

11.9 SBDとフェンシングのテスト

SBDがノードフェンシング目的で期待どおりに機能するかどうかをテストするには、次のいずれかまたはすべての方法を使用します。

ノードのフェンシングを手動でトリガする

ノード `NODENAME` のフェンシングアクションをトリガするには:

```
root # crm node fence NODENAME
```

当該ノードがフェンシングされているかどうか、および `stonith-watchdog-timeout` の時間が経過した後に他のノードが当該ノードをフェンシングされたとしてみなしているかどうかを確認します。

SBD障害のシミュレーション

1. SBD inquisitorのプロセスIDを特定します。

```

root # systemctl status sbd
• sbd.service - Shared-storage based fencing daemon

   Loaded: loaded (/usr/lib/systemd/system/sbd.service; enabled; vendor preset:
disabled)
   Active: active (running) since Tue 2018-04-17 15:24:51 CEST; 6 days ago
     Docs: man:sbd(8)
  Process: 1844 ExecStart=/usr/sbin/sbd $SBD_OPTS -p /var/run/sbd.pid watch
(code=exited, status=0/SUCCESS)
 Main PID: 1859 (sbd)
    Tasks: 4 (limit: 4915)
   CGroup: /system.slice/sbd.service
           └─1859 sbd: inquisitor

[...]

```

2. SBD inquisitorプロセスを終了することにより、SBD障害をシミュレーションします。この例では、SBD inquisitorのプロセスIDは 1859 です。

```

root # kill -9 1859

```

当該ノードは積極的にセルフフェンスを行います。他のノードは、当該ノードの喪失を認識し、stonith-watchdog-timeout の時間経過後に当該ノードがセルフフェンスを行ったとみなします。

監視動作の障害によるフェンシングのトリガ

通常の設定では、リソース停止動作の障害によって、フェンシングがトリガされます。フェンシングを手動でトリガするために、リソース停止動作の障害を発生させることができます。あるいは、以下に説明するように、リソース監視動作の設定を一時的に変更して、監視障害を発生させることができます。

1. リソース監視動作の on-fail=fence プロパティを設定します。

```

op monitor interval=10 on-fail=fence

```

2. 監視動作の障害を発生させます(たとえば、リソースがサービスに関連する場合は、それぞれのデーモンを終了させます)。
この障害により、フェンシングアクションがトリガされます。

11.10 ストレージ保護のための追加メカニズム

STONITHによるノードフェンシング以外に、リソースレベルでストレージ保護を実現する他の方法があります。たとえば、SCSI-3とSCSI-4は永続予約を使用しますが、sfex はロック機構を提供します。両方の方法について以下のサブセクションで説明します。

11.10.1 sg_persistリソースの設定

SCSI仕様3および4では、「永続予約」が定義されています。これらはSCSIプロトコル機能であり、I/Oフェンシングとフェールオーバーに使用できます。この機能は、`sg_persist` Linuxコマンドで実装されます。



注記: SCSIディスクの互換性

`sg_persist` のバッキングディスクは、SCSIディスクとの互換性が必要です。`sg_persist` は、SCSIディスクやiSCSI LUNなどのデバイスでのみ機能します。IDE、SATA、またはSCSIプロトコルをサポートしないブロックデバイスでは、使用しないでください。

続行する前に、お使いのディスクが永続予約をサポートしているかどうかを確認してください。次のコマンドを使用します(`DISK`をデバイス名で置き換えてください)。

```
root # sg_persist -n --in --read-reservation -d /dev/DISK
```

結果に、ディスクが永続予約をサポートしているかどうかが表示されます。

- サポートされているディスク:

```
PR generation=0x0, there is NO reservation held
```

- サポートされていないディスク:

```
PR in (Read reservation): command not supported  
Illegal request, Invalid opcode
```

上記のようなエラーメッセージが表示された場合は、古いディスクをSCSIと互換性のあるディスクに交換してください。それ以外の場合は、以下の手順に従います。

1. プリミティブリソース `sg_persist` を作成するには、`root` として次のコマンドを実行します。

```
root # crm configure  
crm(live)configure# primitive sg sg_persist \  
    params devs="/dev/sdc" reservation_type=3 \  
    op monitor interval=60 timeout=60
```

2. `sg_persist` プリミティブをマスタ-スレーブグループに追加します。

```
crm(live)configure# ms ms-sg sg \  
    meta master-max=1 notify=true
```

3. いくつかのテストをします。リソースがマスタ/スレーブ構成のステータスにある場合、マスタインスタンスが実行されているクラスタノードでは `/dev/sdc1` をマウントして書き込みを行えますが、スレーブインスタンスが実行されているクラスタノードでは書き込みが禁止されます。

4. Ext4のファイルシステムプリミティブを追加します。

```
crm(live)configure# primitive ext4 ocf:heartbeat:Filesystem \
    params device="/dev/sdc1" directory="/mnt/ext4" fstype=ext4
```

5. `sg_persist` マスタとファイルシステムリソースの間に、次の順序関係とコロケーションを追加します。

```
crm(live)configure# order o-ms-sg-before-ext4 inf: ms-sg:promote ext4:start
crm(live)configure# colocation col-ext4-with-sg-persist inf: ext4 ms-sg:Master
```

6. `show` コマンドで、すべての変更内容を確認します。

7. 変更をコミットします。

詳細については、`sg_persist` のマニュアルページを参照してください。

11.10.2 sfexを使用した排他的なストレージアクティブ化の保証

このセクションでは、共有ストレージへのアクセスを1つのノードに排他的にロックする低レベルの追加メカニズムである `sfex` を紹介します。ただし、`sfex` は、STONITHと置き換えることはできないので注意してください。`sfex` には共有ストレージが必要なので、上記で説明したSBDノードフェンシングメカニズムは、ストレージの別のパーティションでを使用することをお勧めします。

設計上、`sfex` は、同時実行が必要なワークロード(OCFS2など)では使用できません。これは、従来のフェールオーバースタイルのワークロードに対する保護の層として機能します。これは、実際にはSCSI-2予約と似ていますが、もっと一般的です。

11.10.2.1 概要

共有ストレージ環境では、ストレージの小さなパーティションが1つ以上のロックの保存用に確保されます。

ノードは、保護されたリソースを取得する前に、まず、保護ロックを取得する必要があります。順序は、Pacemakerによって強制されます。`sfex` コンポーネントは、Pacemakerがスプリットブレイン条件に制約されても、ロックが2回以上付与されないことを保証します。

ノードのダウンが永続的にロックをブロックせず、他のノードが続行できるように、これらのロックも定期的に更新される必要があります。

11.10.2.2 設定

次に、sfexで使用する共有パーティションの作成方法と、CIBでsfexロック用にリソースを設定する方法を説明します。1つのsfexパーティションは任意の数のロックを保持でき、ロックごとに1KBのストレージスペースを割り当てする必要があります。デフォルトでは、`sfex_init` はパーティション上にロックを1つ作成します。

！ 重要: 要件

- sfex用の共有パーティションは、保護するデータと同じ論理ユニットにある必要があります。
- 共有されたsfexパーティションは、ホストベースのRAIDやDRBDを使用してはなりません。
- LVM2論理ボリュームを使用することは可能です。

手順 11.9: SFEXパーティションを作成する

1. sfexで使用する共有パーティションを作成します。このパーティションの名前を書き留め、以降の手順の `/dev/sfex` をこの名前で置き換えます。
2. 次のコマンドでsfexメタデータを作成します。

```
root # sfex_init -n 1 /dev/sfex
```

3. メタデータが正しく作成されたかどうか検証します。

```
root # sfex_stat -i 1 /dev/sfex ; echo $?
```

現在、ロックがかかっていないので、このコマンドは、2を返すはずです。

手順 11.10: SFEXロック用リソースを設定する

1. sfexロックは、CIB内のリソースを介して表現され、次のように設定されます。

```
crm(live)configure# primitive sfex_1 ocf:heartbeat:sfex \  
# params device="/dev/sfex" index="1" collision_timeout="1" \  
    lock_timeout="70" monitor_interval="10" \  
# op monitor interval="10s" timeout="30s" on-fail="fence"
```

2. sfexロックによってリソースを保護するには、保護対象のリソースとsfexリソース間の必須の順序付けと配置の制約を作成します。保護対象のリソースが `filesystem1` というIDを持つ場合は、次のようになります。

```
crm(live)configure# order order-sfex-1 inf: sfex_1 filesystem1
crm(live)configure# colocation col-sfex-1 inf: filesystem1 sfex_1
```

3. グループ構文を使用する場合は、sfexリソースを最初のリソースとしてグループに追加します。

```
crm(live)configure# group LAMP sfex_1 filesystem1 apache ipaddr
```

11.11 その他の情報

- `man sbd`
- <https://github.com/ClusterLabs/sbd> 

12 アクセス制御リスト

crmシェル(crmsh)またはHawk2などのクラスタ管理ツールは、root ユーザまたは haclient グループ内のユーザが使用できます。デフォルトで、これらのユーザは完全な読み込み/書き込みのアクセス権を持ちます。アクセスを制限するか、または詳細なアクセス権を割り当てるには、「アクセス制御リスト」(ACL)を使用できます。

アクセス制御リストは、順序付けされたアクセスルールセットで構成されています。各ルールにより、クラスタ設定の一部への読み込みまたは書き込みアクセスの許可、またはアクセスの拒否が行われます。ルールは通常、組み合わせて特定の役割を生成し、ユーザを自分のタスクに一致する役割に割り当てることができます。



注記: CIB構文検証バージョンとACLとの違い

このACLマニュアルは、pacemaker-2.0 以上のCIB構文バージョンでCIBを検証する場合にのみ適用します。この検証方法およびCIBバージョンのアップグレード方法の詳細については、[注記: CIB構文バージョンのアップグレード](#)を参照してください。

SUSE Linux Enterprise High Availability Extension 11 SPxからアップグレードする一方で、それまで使用してきたCIBバージョンを保持する場合は、SUSE Linux Enterprise High Availability Extension 11 SP3以前の『管理ガイド』で「アクセス制御リスト」の章を参照してください。<https://documentation.suse.com/sle-ha-11> から入手できます。

12.1 要件と前提条件

クラスタでACLの使用を開始する前に、次の条件が満たされていることを確認します。

- NIS、Active Directoryを使用するか、またはすべてのノードに同じユーザを手動で追加して、クラスタ内のすべてのノード上に同じユーザがいることを確認します。
- ACLでアクセス権を変更したいすべてのユーザが haclient グループに属している必要があります。
- すべてのユーザが絶対パス /usr/sbin/crm でcrmshを実行する必要があります。
- 権限のないユーザがcrmshを実行する場合は、/usr/sbinを使用して、PATH 変数を展開する必要があります。

！ 重要: デフォルトのアクセス権

- ACLはオプションの機能です。デフォルトでは、ACLの使用は無効になっています。
- ACL機能が無効化された場合、root および haclient グループに属するすべてのユーザは、クラスタ設定への完全な読み込み/書き込みアクセス権を持ちます。
- ACLが有効化され、設定される場合でも、root およびデフォルトのCRM所有者 haclient は両方とも、「常に」クラスタ設定への完全なアクセス権を持ちます。

ACLを使用するには、XPathに関するいくつかの知識が必要になります。XPathはXMLドキュメントでノードを選択するための言語です。<http://en.wikipedia.org/wiki/XPath> を参照するか、<http://www.w3.org/TR/xpath/> の仕様を確認してください。

12.2 クラスタでのACLの使用の有効化

ACLの設定を開始する前に、ACLの使用を「有効にする」必要があります。有効にするには、`crmsd`で次のコマンドを使用します。

```
root # crm configure property enable-acl=true
```

または、手順12.1「Hawk2でのACLの使用の有効化」で説明するように、Hawk2を使用します。

手順 12.1: HAWK2でのACLの使用の有効化

1. Hawk2にログインします。

```
https://HAWKSERVER:7630/
```

2. 左のナビゲーションバーで、[クラスタ設定]を選択して、グローバルクラスタオプションとそれらの現在の値を表示します。
3. [クラスタ設定]の下にある空のドロップダウンボックスをクリックし、[enable-acl]を選択してパラメータを追加します。デフォルト値 No で追加されます。
4. 値を Yes に設定して変更を適用します。

12.3 ACLの基¥'96¥'7b事項

アクセス制御リストは、順序付けされたアクセスルールセットで構成されています。各ルールにより、クラスタ設定の一部への読み込みまたは書き込みアクセスの許可、またはアクセスの拒否が行われます。ルールは通常、組み合わせて特定の役割を生成し、ユーザを自分のタスクに一致する役割に割り当てることができます。ACLの役割はCIBへのアクセス権を表すルールのセットです。ルールは次の要素で構成されています。

- read、write、または deny のようなアクセス権。
- ルールを適用する場所の指定。種類、ID参照、またはXPath式を使用して指定できます。

通常、ACLを役割にバンドルし、システムユーザ(ACLターゲット)に特定の役割を割り当てると便利です。ACLルールを作成するためには、次の2つの方法があります。

- 12.3.1項「XPath式によるACLルールの設定」。ACLルールを作成するためには、その記述言語であるXMLの構造を理解する必要があります。
- 12.3.2項「短縮によるACLルールの設定」。簡略構文を作成し、ACLルールが一致するオブジェクトに適用します。

12.3.1 XPath式によるACLルールの設定

XPathによってACLルールを管理するには、その記述言語であるXMLの構造を理解する必要があります。XMLでクラスタ設定を表示する次のコマンドで構造を取得します(例 12.1を参照)。

```
root # crm configure show xml
```

例 12.1: XML内のクラスタ設定の例

```
<num_updates="59"
  dc-uuid="175704363"
  crm_feature_set="3.0.9"
  validate-with="pacemaker-2.0"
  epoch="96"
  admin_epoch="0"
  cib-last-written="Fri Aug 8 13:47:28 2014"
  have-quorum="1">
<configuration>
  <crm_config>
    <cluster_property_set id="cib-bootstrap-options">
      <nvpair name="stonith-enabled" value="true" id="cib-bootstrap-options-stonith-
enabled"/>
```

```

    [...]
    </cluster_property_set>
</crm_config>
<nodes>
  <node id="175704363" uname="alice"/>
  <node id="175704619" uname="bob"/>
</nodes>
<resources> [...] </resources>
<constraints/>
<rsc_defaults> [...] </rsc_defaults>
<op_defaults> [...] </op_defaults>
<configuration>
</cib>

```

XPath言語を使用して、このXMLドキュメント内のノードを見つけることができます。たとえば、ルートノード(`cib`)を選択するには、XPath式 `/cib` を使用します。グローバルクラスタ設定を見つけるには、XPath式 `/cib/configuration/crm_config` を使用します。

一例として、表12.1「オペレータ役割 - アクセスタイプおよびXPath式」は、「オペレータ」の役割を作成するためのパラメータ(アクセスタイプおよびXPath式)を示しています。この役割を持つユーザは、2番目の列で説明されるタスクのみ実行することができ、リソースを再構成することはできません(たとえば、パラメータや操作の変更など)。また、コロケーションや順序の制約の設定を変更することもできません。

表 12.1: オペレータ役割 - アクセスタイプおよびXPATh式

タイプ	XPath/説明
書き込み	<pre>//crm_config//nvpair[@name='maintenance-mode']</pre> <p>クラスタ保守モードをオンまたはオフにします。</p>
書き込み	<pre>//op_defaults//nvpair[@name='record-pending']</pre> <p>保留中の操作を記録するかを選択します。</p>
書き込み	<pre>//nodes/node//nvpair[@name='standby']</pre> <p>ノードをオンラインまたはスタンバイモードで設定します。</p>
書き込み	<pre>//resources//nvpair[@name='target-role']</pre> <p>リソースを開始、停止、昇格または降格します。</p>
書き込み	<pre>//resources//nvpair[@name='maintenance']</pre>

タイプ	XPath/説明
	リソースを保守モードにするかどうかを選択します。
書き込み	<pre>//constraints/rsc_location</pre> <p>リソースをノードから別のノードにマイグレート/移動します。</p>
読み込み	<pre>/cib</pre> <p>クラスタのステータスを表示します。</p>

12.3.2 短縮によるACLルールの設定

XML構造を扱いたくないユーザ向けには、より簡単な方法があります。

たとえば、次のXPathを検討します。

```
//*[@id="rsc1"]
```

このXPathは、IDが `rsc1` であるXMLノードをすべて探し出します。

短縮構文はこのように書かれます。

```
ref:"rsc1"
```

これは制約にも使用できます。これが冗長なXPathです。

```
//constraints/rsc_location
```

短縮構文はこのように書かれます。

```
type:"rsc_location"
```

短縮構文は `crmsh` および `Hawk2` で使用できます。CIBデーモンは一致するオブジェクトにACLルールを適用する方法を認識しています。

12.4 Hawk2によるACLの設定

次の手順は、`monitor` 役割を定義し、それをユーザに割り当てることで、クラスタ設定への読み込み専用アクセスを設定する方法を示しています。または、[手順12.4「監視の役割を追加して、crmshを持つユーザに割り当てる」](#)で説明されているように、`crmsh` を使用してこの操作を実行することもできます。

手順 12.2: HAWK2によるMONITOR役割の追加

1. Hawk2にログインします。

`https://HAWKSERVER:7630/`

2. 左のナビゲーションバーで、[役割]を選択します。
3. [作成]をクリックします。
4. 固有な[役割ID]として、monitor などを入力します。
5. アクセス[権利]として、Readを選択します。
6. [Xpath]として、XPath式 /cib を入力します。

SUSE Hawk クラスタ詳細の表示

管理
ステータス
ダッシュボード
履歴

設定
リソースの追加
制約の追加
ウィザード
設定の編集
クラスタ設定
コマンドログ

アクセス制御
役割
ターゲット

Copyright © 2009-2017 SUSE, LLC

役割の作成

役割ID:

ルール:

権利:

XPath:

オブジェクトタイプ:

参照:

ACLの役割

ACLの役割は、CIBへのアクセス権を表すルールセットです。
各ルールは次の要素で構成されます。

- アクセス権 (読み込み、書き込み、または拒否)
- ルールを適用する場所の指定 (XPath式、タイプ、またはID参照)

役割の作成

役割ID: 固有のIDを定義します。

権利: アクセス権 (読み込み / 書き込み / 拒否) を選択します。

Xpath: アクセス権を適用するCIB要素のXPath式を入力します
(例: 場所の制約に適用する場合は `//constraints/rsc_location`)。

タイプ: アクセス権を適用するCIB XML要素の名前を入力します
(例: 場所の制約に適用する場合は `rsc_location`)。

参照: アクセス権を適用するCIB XML要素のIDを入力します
(例: ID `rsc1` を持つすべてのXML要素に適用する場合は、`rsc1`)。

7. [作成]をクリックします。
この操作は、monitor の名前を持つ新しい役割を作成して、read の権利を設定し、XPath式 /cib を使用してCIB内のすべての要素に適用します。
8. 必要に応じてプラスアイコンをクリックしてルールを追加し、個別のパラメータを指定します。
9. 上矢印や下矢印のボタンを使用して、個別のルールをソートできます。

手順 12.3: HAWK2によるターゲットの役割割当

手順 12.2で作成した役割をシステムユーザ(ターゲット)に割り当てるには、次の手順に従います。

1. Hawk2にログインします。

2. 左のナビゲーションバーで、[ターゲット]を選択します。
3. システムユーザ(ACLターゲット)を作成するには、[作成]をクリックして、固有の[ターゲットID]を入力します(例: `tux`)。このユーザが `haclient` グループに属することを確認します。
4. ターゲットに役割を割り当てるには、1つ以上の[役割]を選択します。
例では、手順 12.2で作成した `monitor` 役割を選択します。



5. 選択内容を確認します。

リソースや制約に対するアクセス権を設定するには、12.3.2項「短縮によるACLルールの設定」で説明したように、短縮構文も使用できます。

12.5 crmshによるACLの設定

次の手順は、`monitor` 役割を定義し、それをユーザに割り当てることで、クラスタ設定への読み込み専用アクセスを設定する方法を示しています。

手順 12.4: 監視の役割を追加して、CRMSHを持つユーザに割り当てる

1. `root` としてログインします。
2. `crmsh`の対話モードを開始します。

```
root # crm configure
crm(live)configure#
```

3. ACLの役割を次のとおり定義します。

- a. `role` コマンドを使用して、新しい役割を定義します。

```
crm(live)configure# role monitor read xpath: "/cib"
```

前のコマンドは、`monitor` の名前を持つ新しい役割を作成して、`read` の権利を設定し、XPath式 `/cib` を使用してCIB内のすべての要素に適用します。必要な場合は、アクセス権およびXPath引数をさらに追加できます。

- b. 必要に応じてさらに役割を追加します。

4. 役割を1つ以上のACLターゲットに割り当てます。このACLターゲットは、該当のシステムユーザです。これらのシステムユーザが `haclient` グループに属していることを確認します。

```
crm(live)configure# acl_target tux monitor
```

5. 変更を確認します:

```
crm(live)configure# show
```

6. 変更をコミットします:

```
crm(live)configure# commit
```

リソースや制約に対するアクセス権を設定するには、12.3.2項「短縮によるACLルールの設定」で説明したように、短縮構文も使用できます。

13 ネットワークデバイスボンディング

多くのシステムで、通常のEthernetデバイスの標準のデータセキュリティ/可用性の要件を超えるネットワーク接続の実装が望ましいことがあります。その場合、数台のEthernetデバイスを集めて1つのボンディングデバイスを設定できます。

ボンディングデバイスの設定には、ボンディングモジュールオプションを使用します。ボンディングデバイスの振る舞いは、ボンディングデバイスのモードによって決定されます。デフォルトの動作は、`mode=active-backup` であり、アクティブなスレーブに障害が発生すると、別のスレーブデバイスがアクティブになります。

Corosyncの使用時は、クラスタソフトウェアでボンディングデバイスが管理されることはありません。したがって、ボンディングデバイスにアクセスする可能性のあるクラスタノードごとに、ボンディングデバイスを設定する必要があります。

13.1 YaSTによるボンディングデバイスの設定

ボンディングデバイスを設定するには、1つのボンディングデバイスに集めることができる数台のEthernetデバイスが必要です。次の手順に従います。

1. `root` としてYaSTを開始し、[システム] > [ネットワーク設定] の順に選択します。
2. [ネットワーク設定] で、[概要] タブに切り替えて、使用可能なデバイスを表示します。
3. 1つのボンディングデバイスに集めるEthernetデバイスにIPアドレスが割り当てられているかどうかチェックします。割り当てられている場合は、それを変更します。
 - a. 選択するデバイスを選択して、[編集] をクリックします。
 - b. 開いている[ネットワークカードのセットアップ] ダイアログの[アドレス] タブで、[リンクとIPなしのセットアップ(ボンディングスレーブ)] オプションを選択します。

ネットワークカードの設定

一般(G) アドレス(A) ハードウェア(R)

デバイスの型(V) 環境設定名(F)

イーサネット eth1

☒ リンクおよび IP の設定無し (ボンディングスレーブ)(K) ☐ iBFTの値を使用する(U)

☐ 可変 IP アドレス(Y) DHCP バージョン 4 と 6 の両方での DHCP

☐ 静的割り当てIPアドレス

IPアドレス(I) サブネットマスク(S) ホスト名(Q)

追加アドレス

IPv4アドレスラベル	IPアドレス	ネットマスク

追加(D) 編集(T) 削除(L)

ヘルプ(H) キャンセル(C) 戻る(B) 次へ(N)

c. [次へ]をクリックして、[ネットワーク設定]ダイアログの[Overview]タブに戻ります。

4. 新しいボンディングデバイスを追加するには:

- a. [追加]をクリックして、[デバイスの型]を[ボンド]に変更します。[次へ]で続行します。
- b. IPアドレスをボンディングデバイスに割り当てる方法を選択します。3つの方法から選択できます。
 - リンクとIPなしのセットアップ(ボンディングスレーブ)
 - 可変IPアドレス(DHCPまたはZeroconf)
 - 固定IPアドレス

ご使用の環境に適合する方法を使用します。Corosyncで仮想IPアドレスを管理する場合は、[静的割り当てIPアドレス]を選択し、インタフェースにIPアドレスを割り当てます。

- c. [ボンドスレーブ]タブに切り替えます。
- d. **ステップ 3.b**でボンディングスレーブとして設定したEthernetデバイスが表示されます。ボンドに含めるEthernetデバイスを選択するには、[ボンドスレーブと順序]の下にある、各デバイスの前のチェックボックスを有効にします。

ネットワークカードの設定

一般(G) アドレス(A) ハードウェア(R) ボンドスレーブ(Q)

ボンドスレーブと順序

☒ eth1 - Ethernet Card 1 設定済み
☒ eth2 - Ethernet Card 2 設定済み
☐ eth3 - Ethernet Card 3 設定済み

上へ(U) 下へ(D)

ボンドドライバオプション(I)

mode=active-backup miimon=100

ヘルプ(H) キャンセル(C) 戻る(B) 次へ(N)

e. [ボンドドライバオプション]を編集します。次のモードを使用できます。

balance-rr

パケットが正しい順序で転送されなくなる代わりに、負荷分散と耐障害性が提供されます。これは、TCPの再構築時などに遅延の原因になる場合があります。

active-backup

耐障害性を提供します。

balance-xor

負荷分散と耐障害性を提供します。

ブロードキャスト

耐障害性を提供します。

802.3ad

接続されるスイッチでサポートされる場合は、ダイナミックリンク集合を提供します。

balance-tlb

発信トラフィックの負荷分散を提供します。

balance-alb

使用中にハードウェアアドレスの変更が可能なネットワークデバイスを使用する場合は、着信トラフィックと発信トラフィックの負荷分散を提供します。

- f. [ボンドドライバオプション]には、パラメータ miimon=100 を必ず追加します。このパラメータがなければ、リンクが定期的にチェックされないため、ボンディングドライバは、障害リンクで引き続きパケットを失う可能性があります。
5. [次へ]をクリックして、[OK]でYaSTを終了し、ボンディングデバイスの設定を完了します。YaSTが /etc/sysconfig/network/ifcfg-bondDEVICENUMBER に設定を書き込みます。

13.2 ボンディングスレーブのホットプラグ

ボンディングスレーブのインタフェースを別のものに置き換える必要が生じることがあります。たとえば、それぞれのネットワークデバイスに常に障害が発生する場合などです。解決方法として、ボンディングスレーブのホットプラグを設定します。デバイスをMACアドレスではなくバスIDによって一致させるために、udev ルールを変更する必要もあります。これにより、ハードウェアが許可する場合は、不具合のあるハードウェア(同じスロットにあるのにMACアドレスが異なるネットワークカードなど)を置き換えることができます。

手順 13.1: YASTによるボンディングスレーブのホットプラグの設定

手動の設定を行う場合は、『SUSE Linux Enterprise High Availability Extension Administration Guide』、「Basic Networking」の章の「Hotplugging of Bonding Slaves」というセクションを参照してください。

1. root としてYaSTを開始し、[システム] > [ネットワーク設定]の順に選択します。
2. [ネットワーク設定]で、[Overview]タブに切り替えて、すでに設定済みのデバイスを表示します。ボンディングスレーブがすでに設定済みの場合、[メモ]列にそのことが示されます。

ネットワークの設定

グローバルオプション(G) 概要(V) ホスト名/DNS ルーティング(U)

名前	IPアドレス	デバイス	メモ
ボンドネットワーク		bond0	
Ethernet Card 0	10.161.10.176	eth0	
Ethernet Card 1	なし	eth1	bond0 でスレーブ化
Ethernet Card 2	なし	eth2	bond0 でスレーブ化
Ethernet Card 3	まだ設定されていません		

Ethernet Card 1 (接続されていません)
 MAC : 52:54:00:58:95:45
 BusID : virtio6

- デバイス名: eth1
- 起動時に自動的に開始する
- ボンディングマスター: bond0

追加(A) 編集(E) 削除(D)

ヘルプ(H) キャンセル(C) OK(O)

3. 1つのボンディングデバイスに集められたEthernetデバイスのそれぞれに対して、次の手順を実行します。
 - a. 選択するデバイスを選択して、[編集]をクリックします。[Network Card Setup]ダイアログが開きます。
 - b. [一般]タブに切り替えて、[デバイスのアクティブ化]が[ホットプラグ]に設定されていることを確認します。
 - c. [ハードウェア]タブに切り替えます。
 - d. [Udevルール]で、[変更]をクリックして[BusID]オプションを選択します。
 - e. [OK]および[次へ]をクリックして、[ネットワーク設定]ダイアログの[Overview]タブに戻ります。この時点でEthernetデバイスエントリをクリックすると、下のペインにバスIDを含むデバイスの詳細が表示されます。
4. [OK]をクリックして変更を確定し、ネットワーク設定を終了します。

ブート時にネットワークセットアップはホットプラグスレーブを待機しませんが、ボンドの準備が整うのを待機します。これには少なくとも1つのスレーブが利用可能であることが必要です。スレーブインタフェースの1つがシステムから削除されると(NICドライバからアンバインド、NICドライバの `rmmod`、または実際のPCIホットプラグ取り外し)、カーネルによってボンドから自動的に削除されます。システムに新しいカードが追加されると(スロットのハードウェアが置換されると)、`udev` は、バスベースの永続名規則を適用することで名前を変更し、`ifup` を呼び出します。`ifup` 呼び出しによって、ボンドに自動的に追加されます。

13.3 その他の情報

全モードおよび多数のオプションの詳細については、[Linux Ethernet Bonding Driver HOWTO] に記載されています。これは、`kernel-source` パッケージをインストールした後に参照できる `/usr/src/linux/Documentation/networking/bonding.txt` ファイルの内容です。

High Availabilityセットアップの場合は、そのファイルで説明されている `miimon` および `use_carrier` オプションが特に重要です。

14 負荷バランス

「負荷分散」によって、外部のクライアントからは、サーバのクラスタが1つの大きな高速サーバであるかのようにみえます。この単一サーバのように見えるサーバは、仮想サーバと呼ばれます。このサーバは、着信要求をディスパッチする1つ以上のロードバランサと実際のサービスを実行しているいくつかの実際のサーバで構成されます。High Availability Extensionの負荷分散設定によって、高度にスケーラブルで可用性の高いネットワークサービス(Web、キャッシュ、メール、FTP、メディア、VoIPなど)を構築できます。

14.1 概念の概要

High Availability Extensionは、負荷分散の2つのテクノロジー(Linux仮想サーバ(LVS)およびHAProxy)をサポートしています。これらの主な相違点は、Linux仮想サーバがOSI第4層(トランスポート)でカーネルのネットワーク層を設定するのに対し、HAProxyは第7層(アプリケーション)のユーザスペースで実行されることにあります。このように、Linux仮想サーバは、より少ないリソースで、より高い負荷を処理します。それに対してHAProxyは、トラフィックを調査し、SSL停止を実行して、トラフィックのコンテンツに基づいたディパッチに関する決定を行います。

一方、Linux仮想サーバには、IPVS (IP Virtual Server)およびKTCPPVS (Kernel TCP Virtual Server)という2つの異なるソフトウェアが組み込まれています。IPVSは第4層の負荷分散を提供するのに対し、KTCPPVSは第7層の負荷分散を提供します。

この項では、高可用性と組み合わせた負荷分散について概説してから、Linux仮想サーバとHAProxyについて簡単に説明します。最後に、追加情報を紹介します。

実際のサーバとロードバランサは、高速LANまたは地理的に分散されたWANのいずれでも、相互に接続できます。ロードバランサは、さまざまなサーバに要求をディスパッチします。ロードバランサによって、クラスタの平行サービスが1つのIPアドレス(仮想IPアドレスまたはVIP)上の仮想サービスであるかのようにみえます。要求のディスパッチでは、IP負荷分散技術か、アプリケーションレベル負荷分散技術を使用できます。クラスタ内のノードのトランスペアレントな追加または削除によって、システムのスケーラビリティが達成されます。

ノードまたはサービスの障害検出と仮想サーバシステム全体の適切な再設定によって、常に高い可用性が実現されます。

いくつかの負荷分散戦略があります。ここに、Linux仮想サーバに適した第4層の各戦略を示します。

- **ラウンドロビン:** 最も簡単な戦略は、各接続を異なるアドレスに順番に指定することです。たとえば、DNSサーバは指定のホスト名に対するいくつかのエントリを持つことができます。DNSラウンドロビンでは、DNSサーバは循環しながらそれらのエントリすべてを順番に返します。このように、異なるクライアントは異なるアドレスを表示します。
- **「最良の」サーバの選択:** これにはいくつかのデメリットがありますが、「応答する最初のサーバ」または「負荷の最も少ないサーバ」アプローチで分散を実装できます。
- **サーバあたりの接続数の分散:** ユーザとサーバ間のロードバランサは、複数のサーバ間でユーザ数を分割できます。
- **地理的位置:** 近くのサーバにクライアントをダイレクトすることができます。

ここに、HAProxyに適した第7層の各戦略を示します。

- **URI:** HTTPコンテンツを調査し、この特定のURIに最適なサーバにディスパッチします。
- **URLパラメータ、RDPクッキー:** セッションパラメータ(ポストパラメータの場合もある)、またはRDP(リモートデスクトッププロトコル)セッションクッキーのHTTPコンテンツを調査し、このセッションを提供するサーバにディパッチします。

一部の重複はありますが、HAProxyはLVS/ `ipvsadm` が不十分なシナリオで使用できます(およびその逆もあり)。

- **SSL停止:** フロントエンドロードバランサは、SSL層を処理できます。このため、クラウドノードは、SSLキーにアクセスする必要はなく、ロードバランサのSSLアクセラレータを利用できます。
- **アプリケーションレベル:** HAProxyはアプリケーションレベルで動作するため、コンテンツストリームによって負荷分散の決定に影響を与えることができます。これにより、クッキーや他のフィルタに基づいた永続化が許可されます。

一方、LVS/ `ipvsadm` は、HAProxyで完全に置き換えることはできません。

- LVSは、ロードバランサがインバウンドストリーム内にのみ配置される「ダイレクトルーティング」をサポートし、アウトバウンドトラフィックは直接クライアントにルーティングされます。これにより、非対称環境でのスループットがかなり向上する可能性があります。
- LVSは、(`conntrackd`を介した)ステートフルな接続テーブルレプリケーションをサポートしています。これにより、クライアントおよびサーバに透過なロードバランサのフェールオーバーが可能になります。

14.2 Linux仮想サーバによる負荷分散の設定

以降のセクションでは、主要なLVSのコンポーネントと概念の概要を示します。その後、High Availability ExtensionでのLinux仮想サーバのセットアップ方法について説明します。

14.2.1 Director

LVSの主要コンポーネントは、`ip_vs` (またはIPVS)カーネルコードです。このコードは、Linuxカーネル内でトランスポート層の負荷分散(レイヤ-4スイッチング)を実装します。IPVSコードを含むLinuxカーネルを実行するノードは、ディレクターと呼ばれます。ディレクターで実行されるIPVSコードは、LVSの必須機能です。

クライアントがディレクターに接続すると、着信要求がすべてのクラスターノードに負荷分散されます。つまり、ディレクターは、変更されたルーティングルール(LVSを機能させる)セットを使用して、パケットを実サーバに転送します。たとえば、ディレクターは、接続の送受信端でないと、受信確認を送信しません。ディレクターは、エンドユーザから実サーバ(要求を処理するアプリケーションを実行するホスト)にパケットを転送する特殊なルータとして動作します。

デフォルトでは、IPVSモジュールはカーネルにインストールされている必要はありません。IPVSカーネルモジュールは、`kernel-default` パッケージに含まれています。

14.2.2 ユーザスペースのコントローラとデーモン

`ldirectord` デーモンは、Linux仮想サーバを管理し、負荷分散型仮想サーバのLVSクラスター内の実サーバを監視するユーザスペースデーモンです。設定ファイル `/etc/ha.d/ldirectord.cf` は、仮想サービスとそれらに関連付けられた実サーバを指定し、LVSリダイレクタとしてサーバを設定する方法を `ldirectord` に指示します。このデーモンは、その初期化時にクラスターの仮想サービスを生成します。

`ldirectord` デーモンは、既知のURLを定期的に要求し、応答を確認することにより、実サーバのヘルスを監視します。障害が発生した実サーバは、ロードバランサで使用可能なサーバのリストから削除されます。サービス監視は、ダウンしていたサーバが回復し、再度機能していることを検出すると、そのサーバを使用可能サーバリストに戻します。すべての実サーバがダウンする場合、Webサービスのリダイレクト先にするフォールバックサーバを指定できます。通常、フォールバックサーバは、ローカルホストであり、Webサービスが一時的に使用できないことについて緊急ページを表示します。

`ldirectord` は `ipvsadm` ツール(`ipvsadm` パッケージ)を使用して、Linuxカーネル内の仮想サーバテーブルを操作します。

14.2.3 パケット転送

ディレクターがクライアントから実サーバにパケットを送信する方法は、3つあります。

NAT (Network Address Translation)

着信要求は仮想IPで着信します。宛先のIPアドレスとポートを、選択した実サーバのIPアドレスとポートに変更することで、着信要求は実サーバに転送されます。実サーバはロードバランサに応答を送信し、そのロードバランサが宛先IPアドレスを変更して、応答をクライアントへ転送します。その結果、エンドユーザは予期されたソースから応答を受信します。すべてのトラフィックはロードバランサを通過するので、通常、ロードバランサがクラスタのボトルネックになります。

IPトンネリング(IP-IPカプセル化)

IPトンネリングでは、あるIPアドレスにアドレス指定されたパケットを別のアドレス(別のネットワーク上でも可能)にリダイレクトできます。LVSは、IPトンネルを介して実サーバに要求を送信し(別のIPアドレスにリダイレクト)、実サーバは、独自のルーティングテーブルを使用して、クライアントに直接応答します。クラスタメンバは、さまざまなサブネットに属することができます。

直接ルーティング

エンドユーザからのパケットを、直接、実サーバに転送します。IPパケットは変更されないため、仮想サーバのIPアドレスのトラフィックを受け付けるように、実サーバを設定する必要があります。実サーバからの応答は、直接、クライアントに送信されます。実サーバとロードバランサは、同じ物理ネットワークセグメントに属する必要があります。

14.2.4 スケジューリングアルゴリズム

クライアントから要求された新しい接続に使用する実サーバの決定は、さまざまなアルゴリズムを使用して実装されます。それらは、モジュールとして使用可能であり、特定のニーズに合わせて調整できます。使用可能なモジュールの概要については、[ipvsadm\(8\)](#)のマニュアルページを参照してください。ディレクターは、クライアントから接続要求を受信すると、スケジュールに基づいて実際のサーバをクライアントに割り当てます。スケジューラは、IPVSカーネルコードの一部として、次の新しい接続を取得する実際のサーバを決定します。

Linux仮想サーバのスケジューリングアルゴリズムの詳細については、<http://kb.linuxvirtualserver.org/wiki/IPVS>を参照してください。また、[ipvsadm](#)のマニュアルページで `--scheduler` を検索してください。

関連するHAProxy負荷分散戦略については、<http://www.haproxy.org/download/1.6/doc/configuration.txt>を参照してください。

14.2.5 YaSTによるIP負荷分散の設定

YaST IP負荷分散モジュールを使用して、カーネルベースのIP負荷分散を設定できます。このモジュールは、ldirectordのフロントエンドです。

IP負荷分散ダイアログにアクセスするには、rootとしてYaSTを開始し、[高可用性] > [IP負荷分散]の順に選択します。または、コマンドラインで「yast2 ip1b」と入力して、rootとしてYaSTクラスタモジュールを起動します。

YaSTモジュールは、その設定を `/etc/ha.d/ldirectord.cf` に書き込みます。YaSTモジュール内で使用できるタブは、設定ファイル `/etc/ha.d/ldirectord.cf` の構造、グローバルオプションの定義、および仮想サービス用オプションの定義に対応しています。

設定例とその結果のロードバランサ/実サーバ間のプロセスについては、[例14.1「単純なldirectord設定」](#)を参照してください。



注記: グローバルパラメータと仮想サーバパラメータ

特定のパラメータを仮想サーバセクションとグローバルセクションの両方で指定した場合は、仮想サーバセクションで定義した値が、グローバルセクションで定義した値に優先します。

手順 14.1: グローバルパラメータを設定する

次の手順では、重要なグローバルパラメータの設定方法を示します。個々のパラメータ（および、ここに記載されていないパラメータ）の詳細については、[ヘルプ]をクリックするか、ldirectordのマニュアルページを参照してください。

1. [確認間隔]で、ldirectordが各実サーバに接続していて、それらがまだオンラインかどうか確認する間隔を定義します。
2. [確認タイムアウト]で、最後の確認後に実サーバが応答する期限を設定します。
3. [障害発生回数]では、ldirectordが、何回、実サーバに要求すると、確認が失敗したと見なされるか定義できます。
4. [ネゴシエーションタイムアウト]で、ネゴシエーション確認のタイムアウトを秒単位で定義します。
5. [フォールバック]で、すべての実サーバがダウンした場合にWebサービスのリダイレクト先にするWebサーバのホスト名とIPアドレスを入力します。
6. 実サーバへの接続ステータスがかわったら、システムにアラートを送信させたい場合は、有効な電子メールアドレスを[電子メールアラート]に入力します。
7. [電子メールアラート頻度]で、実サーバにアクセスできない状態が続く場合、何秒後に電子メールアラートを繰り返すか定義します。

8. [電子メールアラートのステータス]で、電子メールアラートを送信する必要のあるサーバのステータスを指定します。複数の状態を定義する場合は、カンマで区切ったリストを使用します。
9. [自動リロード]で、変更の有無について、`ldirectord`に設定ファイルを継続的に監視させるかどうか定義します。yesに設定した場合は、変更のたびに、設定ファイルが自動的にリロードされます。
10. [休止]スイッチで、障害が発生した実サーバをカーネルのLVSテーブルから削除するかどうか定義します。[はい]に設定すると、障害のあるサーバは削除されません。代わりに、それらの重み付けが0に設定され、新しい接続が受け入れられなくなります。すでに確立している接続は、タイムアウトするまで持続します。
11. ロギングに代替パスを使用するには、[ログファイル]でログファイルのパスを指定します。デフォルトでは、`ldirectord`は、そのログファイルを `/var/log/ldirectord.log` に書き込みます。

 IPLB - グローバルな設定

グローバルな設定(G)		仮想サーバの設定(V)	
チェック間隔	チェックタイムアウト	失敗回数	ネゴシエートタイムアウト
<input type="text" value="5"/>	<input type="text" value="3"/>	<input type="text"/>	<input type="text"/>
フォールバック	電子メールアラート		
<input type="text"/>	<input type="text"/>		
コールバック	電子メールアラートの周期		
<input type="text"/>	<input type="text"/>		
エキスパート	電子メールアラートのステータス		
<input type="text"/>	<input type="text"/>		
自動再ロード	休止	フォーク	管理
<input type="text" value="on"/>	<input type="text" value="on"/>	<input type="text" value="on"/>	<input type="text" value="on"/>
ログファイル			
<input type="text"/>			
ヘルプ		キャンセル(Q) OK(O)	

図 14.1: YAST IP負荷分散 - グローバルパラメータ

手順 14.2: 仮想サービスを設定する

仮想サービスごとに、2、3のパラメータを定義することによって、1つ以上の仮想サービスを設定できます。次の手順で、仮想サービスの重要なパラメータを設定する方法を示します。個々のパラメータ(および、ここに記載されていないパラメータ)の詳細については、[ヘルプ]をクリックするか、`ldirectord`のマニュアルページを参照してください。

1. YaST IP負荷分散モジュール内で、[仮想サーバ設定]タブに切り替えます。
2. [追加]で新しい仮想サーバを追加するか、[編集]で既存の仮想サーバを編集します。新しいダイアログに、使用可能なオプションが表示されます。
3. [仮想サーバ]で、共有仮想IPアドレス(IPv4またはIPv6)とポートを入力します。これらのアドレスとポートで、ロードバランサと実サーバをLVSとしてアクセスできます。IPアドレスとポート番号の代わりに、ホスト名とサービスも指定できます。または、ファイアウォールマークを使用することもできます。ファイアウォールマークは、VIP:port サービスの任意の集まりを1つの仮想サービスにまとめる方法です。
4. [実サーバ]で実際のサーバを指定するには、サーバのIPアドレス(IPv4、IPv6、またはホスト名)、ポート(またはサービス名)、および転送方法を入力する必要があります。転送方法は、gate、ipip、またはmasqのいずれかにする必要があります(14.2.3項「[パケット転送](#)」参照)。
[追加]ボタンをクリックし、実サーバごとに必要な引数を入力します。
5. [確認タイプ]で、実サーバがまだアクティブかどうかをテストするために実行する必要がある確認のタイプを選択します。たとえば、要求を送信し、応答に予期どおりの文字列が含まれているかどうか確認するには、[ネゴシエーション]を選択します。
6. [確認のタイプ]を[ネゴシエーション]に設定した場合は、監視するサービスのタイプも定義する必要があります。[サービス]ドロップダウンボックスから選択してください。
7. [要求]で、確認間隔中に各実サーバで要求されるオブジェクトへのURLを入力します。
8. 実サーバからの応答に一定の文字列(「I am alive」メッセージ)が含まれているかどうか確認する場合は、一致する必要がある正規表現を定義します。正規表現を[受信]に入力します。実サーバからの応答にこの表現が含まれている場合、実サーバはアクティブとみなされます。
9. **ステップ 6**で選択した[サービス]のタイプによっては、認証のためのパラメータをさらに指定する必要があります。[認証タイプ]タブに切り替えて、[ログイン]、[パスワード]、[データベース]、または[シークレット]などの詳細を入力します。詳細については、YaSTヘルプのテキストか、ldirectordのマニュアルページを参照してください。
10. [その他]タブに切り替えます。
11. ロードに使用する[スケジューラ]を選択します。使用可能なスケジューラについては、ipvsadm(8)のマニュアルページを参照してください。
12. 使用する[プロトコル]を選択します。仮想サービスをIPアドレスとポートとして指定する場合は、プロトコルをtcpまたはudpのどちらかにする必要があります。仮想サービスをファイアウォールマークとして指定する場合は、プロトコルをfwmにする必要があります。

13. 必要な場合は、さらにパラメータを定義します。**[OK]**を選択して、設定を確認します。YaSTが設定を `/etc/ha.d/ldirectord.cf` に書き込みます。

IPLB - 仮想サーバの環境設定

仮想サーバ

192.168.0.200

実サーバ

192.168.0.110 gate	追加(A)
192.168.0.210 gate	削除(D)
	編集(E)

チェックタイプ(H)
認証タイプ(U)
その他(I)

チェックタイプ	チェックポート	サービス	チェックコマンド
negotiate ▼	<input type="text"/>	<input type="button" value="▽"/>	<input type="text"/>

httpメソッド	要求	受信
<input type="button" value="▽"/>	<input type="text"/>	<input type="text"/>

仮想ホスト	フォールバック
<input type="text"/>	<input type="text" value="127.0.0.1:80"/>

ヘルプ
キャンセル(Q)
OK(O)

図 14.2: YAST IP負荷分散 - 仮想サービス

例 14.1: 単純なLDIRECTORD設定

図14.1「YaST IP負荷分散 - グローバルパラメータ」と図14.2「YaST IP負荷分散 - 仮想サービス」で示された値を使用すると、次のような設定になり、/etc/ha.d/ldirectord.cf で定義されます。

```
autoreload = yes ①
checkinterval = 5 ②
checktimeout = 3 ③
quiescent = yes ④
virtual = 192.168.0.200:80 ⑤
checktype = negotiate ⑥
fallback = 127.0.0.1:80 ⑦
protocol = tcp ⑧
real = 192.168.0.110:80 gate ⑨
real = 192.168.0.120:80 gate ⑨
receive = "still alive" ⑩
request = "test.html" ⑪
scheduler = wlc ⑫
service = http ⑬
```

- ① ldirectordが変更の有無について設定ファイルを継続的に確認するように定義します。
- ② 実サーバがまだオンラインかどうか確認するため、ldirectordが各実サーバに接続する間隔。
- ③ 最後の確認後、実サーバが応答しなければならない時間的な期限
- ④ 障害が発生した実サーバをカーネルのLVSテーブルから削除せず、代わりに、それらのサーバの重み付けを0に設定します。
- ⑤ LVSの仮想IPアドレス(VIP)。LVSはポート80で使用できます。
- ⑥ 実サーバがまだアクティブかどうかをテストするための確認のタイプ。
- ⑦ このサービス用のすべての実サーバがダウンしている場合に、Webサービスのリダイレクト先にするサーバ。
- ⑧ 使用するプロトコル。
- ⑨ ポート80で利用できる2つの実サーバが定義されています。パケットの転送方法がgateなので、直接ルーティングが使用されます。
- ⑩ 実サーバからの応答文字列内で一致する必要がある正規表現。
- ⑪ 確認間隔中に、各実サーバで要求されるオブジェクトへのURI。
- ⑫ 負荷分散に使用するスケジューラが選択されています。
- ⑬ 監視するサービスのタイプ

この設定を使用すると、次のような処理フローになります: ldirectordが、5秒ごとに(②)各実サーバに接続し、⑨と⑪で指定されているように、192.168.0.110:80/test.htmlまたは192.168.0.120:80/test.htmlを要求します。予期された still alive 文字列(⑩)を、最後の確認から 3秒以内(③)に実サーバから受信しない場合は、実サーバが使用可能なサーバのプールから削除されます。ただし、quiescent=yesが設定されているので(④)、実サーバは、LVSテーブルからは削除されません。代わりに、その重み付けが0に設定されます。その結果、この実サーバへの新しい接続は受け付けられなくなります。すでに確立されている接続は、タイムアウトするまで持続します。

14.2.6 追加設定

YaSTによる ldirectord の設定に加えて、LVS設定を完了するには、次の条件を満たす必要があります。

- 実サーバは、必要なサービスを提供するように正しく設定します。
- 負荷分散サーバは、IP転送を使用して実サーバにトラフィックをルーティングできる必要があります。実サーバのネットワーク設定は、選択したパケット転送方法によって左右されます。

- 負荷分散サーバをシステム全体のシングルポイント障害にしないため、ロードバランサのバックアップを1つ以上セットアップする必要があります。クラスタ設定では、ldirectordにプリミティブリソースを設定して、ハードウェア障害の場合にldirectordが他のサーバにフェールオーバーできるようにします。
- ロードバランサのバックアップにも、その作業を達成するために、ldirectord 設定ファイルが必要なので、ロードバランサのバックアップとして使用するすべてのサーバ上で /etc/ha.d/ldirectord.cf が使用できるようにします。設定ファイルは、4.5項「すべてのノードへの設定の転送」で説明されているように、Csync2で同期できます。

14.3 HAProxyによる負荷分散の設定

次のセクションでは、HAProxyの概要とHigh Availabilityでのセットアップ方法について説明します。ロードバランサは、すべての要求をそのバックエンドサーバに分配します。あるマスタで障害が発生すると、スレーブがマスタになることを意味する、アクティブ/パッシブとして設定されます。このようなシナリオでは、ユーザは中断したことに気付きません。

このセクションでは、次のセットアップを使用します。

- ロードバランサー、IPアドレス 192.168.1.99
- 仮想の浮動IPアドレス 192.168.1.99
- サーバ(通常はWebコンテンツ用) www.example1.com (IP: 192.168.1.200)および www.example2.com (IP: 192.168.1.201)

HAProxyを設定するには、次の手順に従います。

1. haproxy パッケージをインストールします。
2. 次のコンテンツを含む /etc/haproxy/haproxy.cfg ファイルを作成します。

```
global ❶
    maxconn 256
    daemon

defaults ❷
    log      global
    mode     http
    option   httplog
    option   dontlognull
    retries  3
    option   redispatch
    maxconn  2000
    timeout  connect    5000 ❸
```

```

timeout client      50s      ④
timeout server      50000    ⑤

frontend LB
  bind 192.168.1.99:80 ⑥
  reqadd X-Forwarded-Proto:\ http
  default_backend LB

backend LB
  mode http
  stats enable
  stats hide-version
  stats uri /stats
  stats realm Haproxy\ Statistics
  stats auth haproxy:password ⑦
  balance roundrobin ⑧
  option httpclose
  option forwardfor
  cookie LB insert
  option httpchk GET /robots.txt HTTP/1.0
  server web1-srv 192.168.1.200:80 cookie web1-srv check
  server web2-srv 192.168.1.201:80 cookie web2-srv check

```

- ① プロセスワイドでOS固有のオプションを含むセクション。

maxconn

プロセスあたりの同時接続の最大数。

デーモン

HAProxyがバックグラウンドで実行する推奨モード。

- ② セクションの宣言後に、他のすべてのセクションのデフォルトパラメータを設定するセクション。次の重要な行があります。

redispatch

接続が失敗した場合にセッションの再ディストリビューションを有効または無効にします。

log

イベントおよびトラフィックのログ記録を有効にします。

mode http

HTTPモードで動作します(HAProxyの推奨モード)このモードでは、サーバへの接続が実行される前に要求が分析されます。RFCに準拠しない要求は拒否されます。

option forwardfor

HTTP X-Forwarded-For ヘッダを要求に追加します。クライアントのIPアドレスを維持する場合は、このオプションが必要です。

- ③ サーバへの接続試行が成功するまで待機する最大時間。
- ④ クライアント側の最大非アクティブ時間。
- ⑤ サーバ側の最大非アクティブ時間。
- ⑥ フロントエンドおよびバックエンドセクションを1つに結合するセクション。

balance leastconn

負荷分散アルゴリズムを定義します。<http://cbonte.github.io/haproxy-dconv/configuration-1.5.html#4-balance> を参照してください。

stats enable ,

stats auth

(stats enable を使用して)統計レポーティングを有効にします。auth オプションは、特定のアカウントに対して認証された統計のログを記録します。

- ⑦ HAProxy Statisticレポートページの認証情報。
- ⑧ 負荷分散はラウンドロビン処理で動作します。

3. 設定ファイルをテストします。

```
root # haproxy -f /etc/haproxy/haproxy.cfg -c
```

4. Csync2の設定ファイル /etc/csync2/csync2.cfg に次の行を追加して、HAProxy設定ファイルが含まれていることを確認します。

```
include /etc/haproxy/haproxy.cfg
```

5. それを同期します。

```
root # csync2 -f /etc/haproxy/haproxy.cfg
root # csync2 -xv
```



注記

Csync2の設定部分は、HAノードが ha-cluster-bootstrap を使用して設定されたことを想定しています。詳細については、『インストールおよびセットアップクイックスタート』を参照してください。

6. Pacemakerによって起動されるため、HAProxyが両方のロードバランサ(aliceおよびbob)で無効になっていることを確認します。

```
root # systemctl disable haproxy
```

7. 新しいCIBを設定します。

```
root # crm configure
crm(live)# cib new haproxy-config
crm(haproxy-config)# primitive haproxy systemd:haproxy \
    op start timeout=120 interval=0 \
    op stop timeout=120 interval=0 \
    op monitor timeout=100 interval=5s \
    meta target-role=Started
crm(haproxy-config)# primitive vip IPAddr2 \
    params ip=192.168.1.99 nic=eth0 cidr_netmask=23 broadcast=192.168.1.255 \
    op monitor interval=5s timeout=120 on-fail=restart \
    meta target-role=Started
crm(haproxy-config)# order haproxy-after-IP Mandatory: vip haproxy
crm(haproxy-config)# colocation haproxy-with-public-IPs inf: haproxy vip
crm(haproxy-config)# group g-haproxy vip haproxy-after-IP
```


8. 新しいCIBを確認し、エラーがあれば修正します。

```
crm(haproxy-config)# verify
```

9. 新しいCIBをコミットします。

```
crm(haproxy-config)# cib use live
crm(live)# cib commit haproxy-config
```

14.4 その他の情報

- <http://www.haproxy.org> 
- <http://www.linuxvirtualserver.org/> にあるプロジェクトのホームページ。
- ldirectordの詳細については、その総合的なマニュアルページを参照してください。
- LVS Knowledge Base: http://kb.linuxvirtualserver.org/wiki/Main_Page .

15 Geoクラスタ(マルチサイトクラスタ)

SUSE® Linux Enterprise High Availability Extension 12 SP5は、ローカルクラスタとメトロエリアクラスタのほかに、地理的に離れたクラスタ(Geoクラスタ。マルチサイトクラスタとも呼ばれます)もサポートしています。これは、それぞれひとつのローカルクラスタで持った複数の地理的に離れたサイトを持てることを意味します。これらクラスタ間のフェールオーバーは、より高いレベルのエンティティである booth によって管理されます。Geo Clustering for SUSE Linux Enterprise High Availability Extensionの個別の拡張として、Geoクラスタに対するサポートが提供されています。Geoクラスタの使用方法和設定方法の詳細については、項目「Geo Clusteringのクイックスタート」または『Geo Clustering Guide』を参照してください。

16 保守タスクの実行

クラスタノードで保守タスクを実行するには、そのノードで実行中のリソースを停止し、それらを移動するか、あるいはそのノードをシャットダウンするか再起動する必要がある場合があります。また、クラスタからリソースの制御を一時的に引き継ぐか、またはリソースを実行中のままにしてクラスタサービスを停止することも必要な場合があります。

この章では、負の影響を及ぼすことなくクラスタノードを手動で切断する方法について説明します。また、クラスタスタックが保守タスクを実行するために提供するさまざまなオプションの概要についても説明します。

16.1 クラスタノードを切断する意味

クラスタノードをシャットダウンまたは再起動する(またはノード上でPacemakerサービスを停止する)場合、次のプロセスがトリガされます。

- ノード上で実行されているリソースは停止されるか、ノードから移動します。
- リソースの停止が失敗するか、タイムアウトする場合、STONITHメカニズムはノードをフェンシングし、シャットダウンします。

手順 16.1: クラスタノードの手動による再起動

ノードをシャットダウンまたは再起動する前に、順序だった方法でノードのサービスをオフにしたい場合は、次の操作を実行します。

1. 再起動またはシャットダウンするノードで、root または同等な権限でログインします。
2. ノードを standby モードにします。

```
root # crm node standby
```

このようにすると、サービスはPacemakerのシャットダウンタイムアウトによって制限されることなく、ノードをオフに移行できます。

3. 以下を使用してクラスタの状態を確認します。

```
root # crm status
```

standby モード状態の各ノードが示されます。

```
[...]
```

```
Node bob: standby
[...]
```

4. そのノードでPacemakerサービスを停止します。

```
root # systemctl stop pacemaker.service
```

5. ノードを再起動します。

ノードが再びクラスタに参加しているかどうかを確認するには:

1. root または同等の権限でノードにログインします。
2. Pacemakerサービスが開始されているかどうかを確認します。

```
root # systemctl status pacemaker.service
```

3. 開始されていない場合は、開始します。

```
root # systemctl start pacemaker.service
```

4. 以下を使用してクラスタの状態を確認します。

```
root # crm status
```

ノードが再びオンラインになっていることが示されます。

16.2 保守タスクのためのさまざまなオプション

Pacemakerはシステム保守を実行するためのさまざまなオプションを提供しています。

クラスタを保守モードにする

グローバルクラスタプロパティ maintenance-mode により、すべてのリソースを瞬時に保守状態にすることができます。クラスタはモニタリングを停止し、ステータスが追跡されなくなります。

ノードを保守モードにする

このオプションにより、特定のノードで実行されているすべてのリソースを瞬時に保守状態にすることができます。クラスタはモニタリングを停止し、ステータスが追跡されなくなります。

ノードをスタンバイモードにする

スタンバイモードのノードはリソースを実行できなくなります。ノード上で実行されているすべてのリソースは移動するか停止されます(他のノードがリソースを実行する資格がない場合)。また、ノード上のすべての監視操作は停止されます(role="Stopped" に設定された操作を除く)。

別のノードで実行されているサービスを提供し続けながら、クラスタ内の1台のノードを停止する必要がある場合は、このオプションを使用できます。

リソースを保守モードにする

リソースに対してこのモードが有効な場合、リソースの監視操作はトリガされません。このリソースで管理されるサービスに手動で介入する必要があり、その間にリソースの監視操作をクラスタに実行させない場合は、このオプションを使用します。

リソースを非管理対象モードにする

`is-managed` メタ属性により、リソースを一時的にクラスタスタックによって管理されている状態から「解放」することができます。これは、このリソースによって管理されるサービスに手動で介入できることを意味します(たとえば、コンポーネントを調整するなど)。ただし、クラスタはリソースの「監視」と障害の報告を継続して行います。クラスタによるリソースの「監視」を停止したい場合は、代わりにリソース単位の保守モードを使用します([リソースを保守モードにする](#)を参照してください)。

16.3 保守作業の準備と終了



警告: データ損失の危険

テストまたは保守作業を実行する必要がある場合は、以下の一般的な手順に従います。

従わない場合、リソースが順序だった方法で起動できない、クラスタノード間でCIBが同期されない、データ損失などの、望ましくない負の影響が及ぼされるリスクがあります。

1. 開始する前に、[16.2項](#)で概説されている、自分の状況に適したオプションを選択します。
2. Hawk2またはcrumshを使用してこのオプションを適用します。
3. 保守タスクまたはテストを実行します。
4. 終了したら、リソース、ノードまたはクラスタを「通常」の操作状態に戻します。

16.4 クラスタを保守モードにする

クラスタをcrmシェル上で保守モードにするには、次のコマンドを使用します。

```
root # crm configure property maintenance-mode=true
```

保守作業が完了した後で、クラスタを通常モードに戻すには次のコマンドを使用します。

```
root # crm configure property maintenance-mode=false
```

手順 16.2: クラスタをHAWK2を使用して保守モードにする

1. 7.2項「ログイン」で説明したように、Webブラウザを起動してクラスタにログインします。
2. 左のナビゲーションバーで、[クラスタ設定]を選択します。
3. [CRMの環境設定]グループで、空のドロップダウンボックスから[maintenance-mode]属性を選択し、プラスアイコンをクリックして追加します。
4. maintenance-mode=trueを設定するには、maintenance-modeの隣のチェックボックスをオンにして、変更を確認します。
5. クラスタ全体の保守作業が完了したら、maintenance-mode属性の隣のチェックボックスをオフにします。
この時点から、High Availability Extensionはクラスタ管理をもう一度引き継ぎます。

16.5 ノードを保守モードにする

crmシェル上でノードを保守モードにするには、次のコマンドを使用します。

```
root # crm node maintenance NODENAME
```

保守作業が完了した後で、ノードを通常モードに戻すには、次のコマンドを使用します。

```
root # crm node ready NODENAME
```

手順 16.3: ノードをHAWK2を使用して保守モードにする

1. 7.2項「ログイン」で説明したように、Webブラウザを起動してクラスタにログインします。
2. 左のナビゲーションバーで、[クラスタステータス]を選択します。
3. 個々のノードのビューのいずれかで、ノードの隣のレンチアイコンをクリックして、[保守]を選択します。
4. 保守タスクが終了したら、ノードの横にあるレンチアイコンをクリックして、[準備完了]を選択します。

16.6 ノードをスタンバイモードにする

ノードをcrmシェル上でスタンバイモードにするには、次のコマンドを使用します。

```
root # crm node standby NODENAME
```

保守作業が完了した後でノードをオンライン状態に戻すには、次のコマンドを使用します。

```
root # crm node online NODENAME
```

手順 16.4: ノードをHAWK2を使用してスタンバイモードにする

1. 7.2項「ログイン」で説明したように、Webブラウザを起動してクラスタにログインします。
2. 左のナビゲーションバーで、[クラスタステータス]を選択します。
3. 個々のノードのビューのいずれかで、ノードの隣のレンチアイコンをクリックして、[スタンバイ]を選択します。
4. ノードの保守タスクを完了します。
5. スタンバイモードを無効化するには、そのノードの隣のレンチアイコンをクリックして[準備完了]を選択します。

16.7 リソースを保守モードにする

crmシェル上でリソースを保守モードにするには、次のコマンドを使用します。

```
root # crm resource maintenance RESOURCE_ID true
```

保守作業が完了した後で、リソースを通常モードに戻すには次のコマンドを使用します。

```
root # crm resource maintenance RESOURCE_ID false
```

手順 16.5: リソースをHAWK2を使用して保守モードにする

1. 7.2項「ログイン」で説明したように、Webブラウザを起動してクラスタにログインします。
2. 左のナビゲーションバーで、[リソース]を選択します。
3. 保守モードまたは非管理対象モードにするリソースを選択し、そのリソースの隣のレンチアイコンをクリックして、[リソースの編集]を選択します。

4. [メタ属性] カテゴリが開きます。
5. 空のドロップダウンボックスから[maintenance]属性を選択し、プラスアイコンをクリックして追加します。
6. maintenance の隣のチェックボックスをオンにして、maintenance属性を yes に設定します。
7. 変更内容を確認します。
8. 該当するリソースの保守作業が完了したら、そのリソースの maintenance 属性の隣のチェックボックスをオフにします。
リソースは、この時点から再びHigh Availability Extensionソフトウェアによって管理されます。

16.8 リソースを非管理対象モードにする

crmシェル上でリソースを非管理対象モードにするには、次のコマンドを使用します。

```
root # crm resource unmanage RESOURCE_ID
```

保守作業が完了した後で再び管理対象モードにするには、次のコマンドを使用します。

```
root # crm resource manage RESOURCE_ID
```

手順 16.6: リソースをHAWK2を使用して非管理対象モードにする

1. 7.2項「ログイン」で説明したように、Webブラウザを起動してクラスタにログインします。
2. 左ナビゲーションバーから、[状態]を選択し、[リソース]リストに移動します。
3. [操作]列で、変更したいリソースの横にある下矢印アイコンをクリックして[編集]を選択します。
リソース設定画面が開きます。
4. [メタ属性]の下で、空のドロップダウンボックスから[is-managed]エントリを選択します。
5. その値を No に設定し、[適用]をクリックします。
6. 保守タスクが終了した後で、[is-managed]を Yes に設定し(デフォルト値です)、変更を適用します。
リソースは、この時点から再びHigh Availability Extensionソフトウェアによって管理されます。

16.9 保守モード中のクラスタノードの再起動



注記: 意味

クラスタまたはノードが保守モードの場合、クラスタリソースを任意に停止したり再起動したりできます。High Availability Extensionはこれらを再起動しようとしません。ノード上のPacemakerサービスを停止する場合、(Pacemakerの管理対象クラスタリソースとして最初に起動された)すべてのデーモンとプロセスの実行は継続されます。

クラスタまたはノードが保守モードのときに、ノード上でPacemakerサービスを起動しようとする場合、Pacemakerはリソースごとに1つのワンショット監視操作(「probe」)を開始し、そのノードで現在どのリソースが実行されているかを評価します。ただし、リソースのステータスを決定する以外の操作は行いません。

クラスタまたはノードが保守モードのときにノードを切断する場合は、次のようにします。

1. 再起動またはシャットダウンするノードで、root または同等な権限でログインします。
2. ocf:pacemaker:controld タイプのリソースを使用しているのか、またはこのタイプのリソースに依存しているリソースを使用しているのかどうかを確認します。ocf:pacemaker:controld タイプのリソースはDLMリソースです。

- a. その場合は、DLMリソースとそれに依存しているすべてのリソースを明示的に停止します。

```
crm(live)resource# stop RESOURCE_ID
```

その理由は、Pacemakerを停止すると、DLMが依存するメンバーシップとメッセージングサービスを持つCorosyncサービスも停止するからです。Corosyncが停止した場合、DLMリソースではスプリットブレインシナリオが発生したと見なされ、フェンシング操作がトリガされます。

- b. そうでない場合は、[ステップ 3](#)に進みます。
3. そのノードでPacemakerサービスを停止します。

```
root # systemctl stop pacemaker.service
```

4. ノードをシャットダウンするか再起動します。

III ストレージおよびデータレプリケーション

- 17 分散ロックマネージャ(DLM:Distributed Lock Manager) 241
- 18 OCFS2 244
- 19 GFS2 253
- 20 DRBD 258
- 21 Cluster Logical Volume Manager(cLVM) 275
- 22 クラスタマルチデバイス(Cluster MD) 286
- 23 Sambaクラスタリング 290
- 24 Relax-and-Recover (Rear)による障害復旧 299

17 分散ロックマネージャ(DLM:Distributed Lock Manager)

カーネル内の分散ロックマネージャ(DLM)は、OCFS2、GFS2、Cluster MD、およびcLVMによって使用されるベースコンポーネントで、各層でアクティブ/アクティブ構成のストレージが提供されます。

17.1 DLM通信の Protokol

シングルポイント障害を回避するには、高可用性クラスタに対する冗長通信パスが重要となります。これはDLM通信においても当てはまります。ネットワークボンディング(Link Aggregation Control Protocol、LACP)が何らかの理由で使用できない場合、Corosyncで冗長通信チャネル(2番目のリング)を定義することを強くお勧めします。詳細については、[手順4.3「冗長通信チャネルの定義」](#)を参照してください。

/etc/corosync/corosync.conf の設定によって、DLMはその通信にTCPプロトコルを使用するか、SCTPプロトコルを使用するかを判断します。

- [rrp_mode]を none に設定する場合(冗長リング設定が無効であることを意味する)、DLMは自動的にTCPを使用します。ただし、冗長通信チャネルを使用しない場合には、TCPリンクがダウンすると、DLM通信は失敗します。
- [rrp_mode]が passive に設定され(これは通常の設定です)、/etc/corosync/corosync.conf の2番目の通信リングが正しく設定されている場合、DLMは自動的にSCTPを使用します。この場合、DLMメッセージングには、SCTPによって提供される冗長性機能があります。

17.2 DLMクラスタリソースの設定

DLMはPacemakerからのクラスタメンバーシップサービスを使用し、それらのサービスはユーザスペースで実行されます。したがって、DLMは、クラスタ内の各ノードに存在するクローンリソースとして設定する必要があります。



注記: いくつかの解決策のためのDLMリソース

OCFS2、GFS2、Cluster MD、およびcLVMのすべてがDLMを使用するため、DLMに1つのリソースを設定することで十分です。DLMリソースはクラスタ内のすべてのノード上で実行されるので、リソースはクローンリソースとして設定されます。

OCFS2およびcLVMの両方を含むセットアップがある場合、OCFS2およびcLVMの両方に1つのDLMリソースを設定することで十分です。

手順 17.1: DLMのベースグループの設定

設定は複数のプリミティブおよび1つのベースクローンを含むベースグループで設定されます。ベースグループとベースクローンはどちらも、後でさまざまなシナリオで使用できます(例: OCFS2およびcLVM)。必要に応じてそれぞれのプリミティブを持つベースグループを拡張する必要があるだけです。ベースグループは内部コロケーションおよび順序付けを持つため、個々のグループ、クローン、その依存性をいくつも指定する必要がなく、セットアップ全体を容易にします。

クラスタ内の1つのノードについて、次の手順を実行してください。

1. シェルを起動し、rootまたは同等のものとしてログインします。
2. `crm configure`を実行します。
3. 次のコマンドを入力して、DLMのプリミティブリソースを作成します。

```
crm(live)configure# primitive dlm ocf:pacemaker:controld \
op monitor interval="60" timeout="60"
```

4. DLMリソースおよび追加のストレージ関連のリソース用にベースグループを作成します。

```
crm(live)configure# group g-storage dlm
```

5. g-storageグループのクローンを作成して、すべてのノードで実行できるようにします。

```
crm(live)configure# clone cl-storage g-storage \
meta interleave=true target-role=Started
```

6. `show`で変更内容をレビューします。
7. すべて正しければ、`commit`で変更を送信し、`exit`でcrmライブ設定を終了します。



注記: STONITHを無効にする際のエラー

STONITHを使用しないクラスタはサポートされていません。テストまたはトラブルシューティングのためにグローバルクラスタオプション `stonith-enabled` を `false` に設定すると、DLMリソースとそれに依存するすべてのサービス(cLVM、GFS2、およびOCFS2など)は起動できません。

18 OCFS2

OCFS 2 (Oracle Cluster File System 2)は、Linux 2.6以降のカーネルに完全に統合されている汎用ジャーナリングファイルシステムです。Oracle Cluster File System 2を利用すれば、アプリケーションバイナリファイル、データファイル、およびデータベースを、共有ストレージ中のデバイスに保管することができます。このファイルシステムには、クラスタ中のすべてのノードが同時に読み書きすることができます。ユーザスペース管理デーモンは、クローンリソースを介して管理され、HAスタック(特に、CorosyncおよびDLM (Distributed Lock Manager))との統合を実現します。

18.1 特長と利点

OCFS2は、たとえば、次のストレージソリューションに使用できます。

- 一般のアプリケーションとワークロード。
- クラスタ中のXENイメージ。Xen仮想マシンと仮想サーバは、クラスタサーバによってマウントされたOCFS2ボリュームに保存できます。これによって、サーバ間でXen仮想マシンを素早く容易に移植できます。
- LAMP (Linux, Apache, MySQL、およびPHP | Perl | Python)スタック。

OCF2は、高パフォーマンスでシンメトリックなパラレルクラスタファイルシステムとして、次の機能をサポートします。

- アプリケーションのファイルを、クラスタ内のすべてのノードで使用できます。ユーザは、クラスタ中のOracle Cluster File System 2ボリュームに1回インストールするだけで構いません。
- すべてのノードが、標準ファイルシステムインタフェースを介して、同時並行的に、ストレージに直接読み書きできるので、クラスタ全体に渡わたって実行されるアプリケーションの管理が容易になります。
- ファイルアクセスがDLMを介して調整されます。ほとんどの場合、DLMによる制御は適切に機能しますが、アプリケーションの設計によっては、アプリケーションとDLMがファイルアクセスの調整で競合すると、スケーラビリティが制限されることがあります。
- すべてのバックエンドストレージで、ストレージのバックアップ機能を利用することができます。共有アプリケーションファイルのイメージを簡単に作成することができるため、災害発生時でも素早くデータを復元することができます。

Oracle Cluster File System 2には、次の機能も用意されています。

- メタデータのキャッシュ処理。
- メタデータのジャーナル処理。
- ノード間にまたがるファイルデータの整合性。
- 最大4KBのマルチブロックサイズ、最大1MBのクラスタサイズ、4PB(ペタバイト)の最大ボリュームサイズをサポートします。
- 32台までのクラスタノードをサポート。
- データベースのパフォーマンスを向上する非同期、直接I/Oのサポート。



注記: OCFS2用サポート

OCFS2は、SUSE Linux Enterprise High Availability Extensionによって提供される、pcmk (Pacemaker)スタックと併用する場合にのみ、SUSEによってサポートされます。o2cbスタックと組み合わせた場合、SUSEはOCFS2をサポートしません。

18.2 OCFS2のパッケージと管理ユーティリティ

SUSE® Linux Enterprise Server 12 SP5上のHigh Availability Extensionには、OCFS2カーネルモジュール(ocfs2)が自動的にインストールされます。OCFS2を使用するには、ocfs2-toolsと、ご使用のカーネルに適合する ocfs2-kmp-* パッケージが、クラスタの各ノードにインストールされていることを確認してください。

ocfs2-tools パッケージには、次に示すOCFS2ボリュームの管理ユーティリティがあります。構文については、各マニュアルページを参照してください。

表 18.1: OCFS2ユーティリティ

OCFS2ユーティリティ	説明
debugfs.ocfs2	デバッグのために、OCFSファイルシステムの状態を調査します。
fsck.ocfs2	ファイルシステムにエラーがないかをチェックし、必要に応じてエラーを修復します。
mkfs.ocfs2	デバイス上にOCFS2ファイルシステムを作成します。通常は、共有物理/論理ディスク上のパーティションに作成します。

OCFS2ユーティリティ	説明
mounted.ocfs2	クラスタシステム上のすべてのOCFS2ボリュームを検出、表示します。OCFS2デバイスをマウントしているシステム上のすべてのノードを検出、表示するか、またはすべてのOCFS2デバイスを表示します。
tunefs.ocfs2	ボリュームラベル、ノードスロット数、すべてのノードスロットのジャーナルサイズ、およびボリュームサイズなど、OCFS2ファイルのシステムパラメータを変更します。

18.3 OCFS2サービスとSTONITHリソースの設定

OCFS2ボリュームを作成する前に、DLMおよびSTONITHリソースをクラスタ内のサービスとして設定する必要があります。

次の手順では、`crm` シェルを使用してクラスタリソースを設定します。[18.6項「Hawk2でのOCFS2リソースの設定」](#)で説明されているように、リソースの設定にはHawk2を使用することもできます。

手順 18.1: STONITHリソースの設定



注記: 必要なSTONITHデバイス

フェンシングデバイスを設定する必要があります。STONITHなしでは、設定内に配置されたメカニズム(`external/sbd` など)は失敗します。

1. シェルを起動し、`root` または同等のものとしてログインします。
2. [手順11.3「SBDデバイスの初期化」](#)で説明されるとおり、SBDパーティションを作成します。
3. `crm configure` を実行します。
4. `external/sdb` をフェンシングデバイスとして設定し、`/dev/sdb2` を共有ストレージ上のハートビートとフェンシング専用のパーティションにします。

```
crm(live)configure# primitive sbd_stonith stonith:external/sbd \
  params pcmk_delay_max=30 meta target-role="Started"
```

5. `show` で変更内容をレビューします。

6. すべて正しければ、`commit` で変更を送信し、`exit` でcrmライブ設定を終了します。

DLMに対するリソースグループの設定の詳細については、[手順17.1「DLMのベースグループの設定」](#)を参照してください。

18.4 OCFS2ボリュームの作成

18.3項「[OCFS2サービスとSTONITHリソースの設定](#)」で説明されているように、DLMクラスタリソースを設定したら、システムがOCFS2を使用できるように設定し、OCFs2ボリュームを作成します。



注記: アプリケーションファイルとデータファイル用のOCFS2ボリューム

一般に、アプリケーションファイルとデータファイルは、異なるOCFS2ボリュームに保存することを推奨します。アプリケーションボリュームとデータボリュームのマウント要件が異なる場合は、必ず、異なるボリュームに保存します。

作業を始める前に、OCFS2ボリュームに使用するブロックデバイスを準備します。デバイスは空き領域のままにしてください。

次に、[手順18.2「OCFS2ボリュームを作成し、フォーマットする」](#)で説明されているように、`mkfs.ocfs2` で、OCFS2ボリュームを作成し、フォーマットします。そのコマンドの重要なパラメータは、[表18.2「重要なOCFS2パラメータ」](#)に一覧されています。詳細情報とコマンド構文については、`mkfs.ocfs2` のマニュアルページを参照してください。

表 18.2: 重要なOCFS2パラメータ

OCFS2パラメータ	説明と推奨設定
ボリュームラベル(<code>-L</code>)	異なるノードへのマウント時に、正しく識別できるように、一意のわかりやすいボリューム名を指定します。ラベルを変更するには、 <code>tune fs.ocfs2</code> ユーティリティを使用します。
クラスタサイズ(<code>-C</code>)	クラスタサイズは、ファイルに割り当てられる、データ保管領域の最小単位です。使用できるオプションと推奨事項については、 <code>mkfs.ocfs2</code> のマニュアルページを参照してください。

OCFS2パラメータ	説明と推奨設定
ノードスロット数(<u>-N</u>)	<p>同時にボリュームをマウントできる最大ノード数を指定します。各ノードについて、OCFS2はジャーナルなどの個別のシステムファイルを作成します。ボリュームにアクセスするノードに、リトルエンディアン形式のノード(AMD64/Intel 64など)とビッグエンディアン形式のノード(System zなど)が混在しても構いません。</p> <p>ノード固有のファイルは、ローカルファイルと呼ばれます。ローカルファイルには、ノードスロット番号が付加されます。たとえば、<code>journal:0000</code> は、スロット番号 <u>0</u> に割り当てられたノードに属します。</p> <p>各ボリュームを同時にマウントすると予期されるノード数に従って、各ボリュームの作成時に、そのボリュームの最大ノードスロット数を設定します。<code>tunefs.ocfs2</code> ユーティリティを使用して、必要に応じてノードスロットの数を増やします。ただし、この値は減らすことはできません。</p> <p><u>-N</u> パラメータを指定しない場合、スロット数はファイルシステムのサイズに基づいて決定されます。</p>
ブロックサイズ(<u>-b</u>)	<p>ファイルシステムがアドレス可能な領域の最小単位を指定します。ブロックサイズは、ボリュームの作成時に指定します。使用できるオプションと推奨事項については、<code>mkfs.ocfs2</code> のマニュアルページを参照してください。</p>
特定機能のオン/オフ(<u>--fs-features</u>)	<p>カンマで区切った機能フラグリストを指定できます。<code>mkfs.ocfs2</code> は、そのリストに従って、それらの機能セットを含むファイルシステムを作成しようとします。機能をオンにするには、その機能をリストに入れます。機能をオフにするには、その名前の前に <u>no</u> を付けます。</p>

OCFS2パラメータ	説明と推奨設定
	使用できるすべてのフラグの概要については、 mkfs.ocfs2 のマニュアルページを参照してください。
事前定義機能(<code>--fs-feature-level</code>)	事前定義されたファイルシステム機能セットから選択できます。使用できるオプションについては、 mkfs.ocfs2 のマニュアルページを参照してください。

[mkfs.ocfs2](#)によるボリュームの作成およびフォーマット時に機能を指定しない場合は、[backup-super](#)、[sparse](#)、[inline-data](#)、[unwritten](#)、[metaecc](#)、[indexed-dirs](#)、および[xattr](#)の各機能がデフォルトで有効になります。

手順 18.2: OCFS2ボリュームを作成し、フォーマットする

クラスタノードの1つだけで、次の手順を実行します。

1. 端末ウィンドウを開いて、[root](#)としてログインします。
2. クラスタがオンラインであることをコマンド `crm status` で確認します。
3. [mkfs.ocfs2](#) ユーティリティを使用して、ボリュームを作成およびフォーマットします。このコマンドの指定形式については、[mkfs.ocfs2](#) マニュアルページを参照してください。
たとえば、最大32台のクラスタノードをサポートする新しいOCFS2ファイルシステムを `/dev/sdb1` 上に作成するには、次のコマンドを使用します。

```
root # mkfs.ocfs2 -N 32 /dev/sdb1
```

18.5 OCFS2ボリュームのマウント

OCFS2ボリュームは、手動でマウントするか、クラスタマネージャでマウントできます(手順18.4「[クラスタリソースマネージャでOCFS2ボリュームをマウントする](#)」参照)。

手順 18.3: OCFS2ボリュームを手動でマウントする

1. 端末ウィンドウを開いて、[root](#)としてログインします。
2. クラスタがオンラインであることをコマンド `crm status` で確認します。
3. コマンドラインから、`mount` コマンドを使ってボリュームをマウントします。



警告: 手動マウントによるOCFS2デバイス

OCFS2ファイルシステムをテスト目的で手動マウントした場合、そのファイルシステムは、いったんマウント解除してから、クラスタリソースで使用してください。

手順 18.4: クラスタリソースマネージャでOCFS2ボリュームをマウントする

High AvailabilityソフトウェアでOCFS2ボリュームをマウントするには、クラスタ内でocfs2ファイルシステムのリソースを設定します。次の手順では、`crm` シェルを使用してクラスタリソースを設定します。18.6項「Hawk2でのOCFS2リソースの設定」で説明されているように、リソースの設定にはHawk2を使用することもできます。

1. シェルを起動し、`root` または同等のものとしてログインします。
2. `crm configure` を実行します。
3. OCFS2ファイルシステムをクラスタ内のすべてのノードにマウントするように、Pacemakerを設定します。

```
crm(live)configure# primitive ocfs2-1 ocf:heartbeat:Filesystem \
  params device="/dev/sdb1" directory="/mnt/shared" \
  fstype="ocfs2" options="acl" \
  op monitor interval="20" timeout="40" \
  op start timeout="60" op stop timeout="60" \
  meta target-role="Stopped"
```

4. `ocfs2-1` プリミティブを手順17.1「DLMのベースグループの設定」で作成した `g-storage` グループに追加します。

```
crm(live)configure# modgroup g-storage add ocfs2-1
```

`add` サブコマンドは、デフォルトで新しいグループメンバーを追加します。ベースグループの内部コロケーションおよび順序付けによって、Pacemakerは、すでに実行している `d1m` リソースも持つノード上で `ocfs2-1` リソースのみ起動します。

5. `show` で変更内容をレビューします。
6. すべて正しいければ、`commit` で変更を送信し、`exit` でcrmライブ設定を終了します。

18.6 Hawk2でのOCFS2リソースの設定

crmシェルを使用して、DLM、およびOCFS2のファイルシステムリソースを手動で設定する代わりに、Hawk2の[セットアップウィザード]のOCFS2テンプレートを使用することもできます。

！ 重要: 手動設定とHawk2との相違点

[セットアップウィザード]のOCFS2テンプレートには、STONITHリソースの設定が含まれません。ウィザードを使用する場合でも、[手順18.1「STONITHリソースの設定」](#)で説明されているように、共有ストレージ上でSBDパーティションを作成し、STONITHリソースを設定する必要があります。

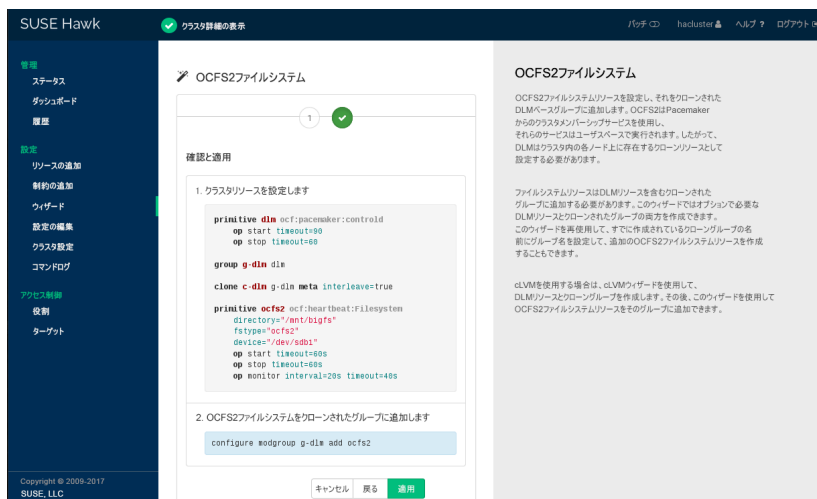
また、Hawk[セットアップウィザード]のOCFS2テンプレートを使用すると、[手順17.1「DLMのベースグループの設定」](#)および[手順18.4「クラスタリソースマネージャでOCFS2ボリュームをマウントする」](#)で説明されている手動設定とは若干異なるリソース設定になります。

手順 18.5: HAWK2の[ウィザード]でのOCFS2リソースの設定

1. Hawk2にログインします。

<https://HAWKSERVER:7630/>

2. 左のナビゲーションバーで、[ウィザード]を選択します。
3. [ファイルシステム]カテゴリを展開して、OCFS2 File Systemを選択します。
4. 画面の指示に従います。オプションについての情報が必要な場合には、オプションをクリックすると、Hawk2は簡単なヘルプテキストを表示します。最後の設定手順が完了したら、[Verify (検証)]を選択して、入力した値を検証します。
CIBに適用する設定スニペットやその他の必要な変更がウィザードに表示されます。



5. 適用予定の変更を確認します。すべてが希望どおりの場合は、変更を適用します。
画面上のメッセージが、アクションに成功したかどうかを示します。

18.7 OCFS2ファイルシステム上でクォータを使用する

OCFS2ファイルシステム上でクォータを使用するには、適切なクォータ機能またはオプションを使用して、ファイルシステムを作成し、マウントします。オプションは `ursquota` (個々のユーザのためのクォータ) または `grpquota` (グループのためのクォータ) です。これらの機能は後ほど、`tunefs.ocfs2` を使用して、マウントされていないファイルシステムで有効にすることもできます。

ファイルシステムで適切なクォータ機能が有効にされている場合、ファイルシステムは、そのメタデータで、各ユーザ(または)グループが使用しているスペースの量とファイルの数を追跡します。OCFS2はクォータ情報をファイルシステムの内部メタデータとして扱うので、`quotacheck` (8) プログラムを実行する必要はありません。すべての機能は `fsck.ocf2`、およびファイルシステムドライバ自体に組み込まれています。

各ユーザまたはグループに課せられている制限の強制を有効にするには、他のファイルシステムの場合と同様に、`quotaon` (8) を実行します。

パフォーマンス上の理由で、各クラスターノードはクォータの計算をローカルに行い、この情報を、10秒ごとに共通の中央ストレージに同期するようになっています。この間隔は、`tunefs.ocfs2` と、`usrquota-sync-interval` および `grpquota-sync-interval` オプションで調整することができます。クォータ情報は必ずしも常に正確というわけではないので、複数のクラスターノードを並列に運用している場合、ユーザまたはグループがクォータ制限をいくらか超えることもあります。

18.8 詳細情報

OCFS2の詳細については、次のリンクを参照してください。

<https://ocfs2.wiki.kernel.org/> 

OCFS2プロジェクトホームページ。

<http://oss.oracle.com/projects/ocfs2/> 

Oracleサイトにある以前のOCFS2プロジェクトのホームページ

<http://oss.oracle.com/projects/ocfs2/documentation> 

プロジェクトの以前のドキュメントホームページ。

19 GFS2

Global File System 2 (GFS2)は、Linuxコンピュータクラスタ用の共有ディスクファイルシステムです。GFS2により、すべてのノードが同じ共有ブロックストレージに直接同時にアクセスすることができます。GFS2には、非接続運用モードがなく、クライアント役割やサーバ役割もありません。GFS2クラスタのすべてのノードがピアとして機能します。GFS2は、クラスタノードを32台までサポートします。クラスタでGFS2を使用する場合は、ハードウェアが共有ストレージへのアクセスを許可し、ロックマネージャがストレージへのアクセスを制御する必要があります。

SUSEでは、パフォーマンスが主要な要件の1つである場合は、クラスタ環境でOCFS2 over GFS2を使用することを推奨しています。弊社のテストは、このような設定では、GFS2と比較してOCFS2の方がパフォーマンスに優れていることを示しています。

19.1 GFS2パッケージおよび管理ユーティリティ

GFS2を使用するには、`gfs2-utils`と、ご使用のカーネルに適合する `gfs2-kmp-*` パッケージが、クラスタの各ノードにインストールされていることを確認してください。

`gfs2-utils` パッケージには、次に示すGFS2ボリュームの管理ユーティリティがあります。構文については、各マニュアルページを参照してください。

表 19.1: GFS2ユーティリティ

GFS2ユーティリティ	説明
<code>fsck.gfs2</code>	ファイルシステムにエラーがないかをチェックし、必要に応じてエラーを修復します。
<code>gfs2_jadd</code>	GFS2ファイルシステムにジャーナルを追加します。
<code>gfs2_grow</code>	GFS2ファイルシステムを拡張します。
<code>mkfs.gfs2</code>	デバイス上にGFS2ファイルシステムを作成します。通常は、共有デバイスまたはパーティションになります。

GFS2ユーティリティ	説明
tunegfs2	次のようなGFS2ファイルシステムパラメータを表示および操作できます(<u>UUID</u> 、 <u>label</u> 、 <u>lockproto</u> および <u>locktable</u>)。

19.2 GFS2サービスとSTONITHリソースの設定

GFS2ボリュームを作成する前に、DLMおよびSTONITHリソースを設定する必要があります。

手順 19.1: STONITHリソースの設定



注記: 必要なSTONITHデバイス

フェンシングデバイスを設定する必要があります。STONITHなしでは、設定内に配置されたメカニズム(external/sbd など)は失敗します。

1. シェルを起動し、root または同等のものとしてログインします。
2. [手順11.3「SBDデバイスの初期化」](#)で説明されるとおり、SBDパーティションを作成します。
3. `crm configure`を実行します。
4. external/sdbをフェンシングデバイスとして設定し、/dev/sdb2を共有ストレージ上のハートビートとフェンシング専用のパーティションにします。

```
crm(live)configure# primitive sbd_stonith stonith:external/sbd \
  params pcmk_delay_max=30 meta target-role="Started"
```

5. `show`で変更内容をレビューします。
6. すべて正しいければ、`commit`で変更を送信し、`exit`でcrmライブ設定を終了します。

DLMに対するリソースグループの設定の詳細については、[手順17.1「DLMのベースグループの設定」](#)を参照してください。

19.3 GFS2ボリュームの作成

19.2項「GFS2サービスとSTONITHリソースの設定」で説明されているように、DLMをクラスタリソースとして設定したら、システムがGFS2を使用できるように設定し、GFS2ボリュームを作成します。



注記: アプリケーションファイルとデータファイル用のGFS2ボリューム

一般に、アプリケーションファイルとデータファイルは、異なるGFS2ボリュームに保存することをお勧めします。アプリケーションボリュームとデータボリュームのマウント要件が異なる場合は、必ず、異なるボリュームに保存します。

作業を始める前に、GFS2ボリュームに使用するブロックデバイスを準備します。デバイスは空き領域のままにしてください。

次に、[手順19.2「GFS2ボリュームの作成とフォーマット」](#)で説明されているよう

に、`mkfs.gfs2` で、GFS2ボリュームを作成し、フォーマットします。そのコマンドの重要なパラメータは、[表19.2「重要なGFS2パラメータ」](#)に一覧されています。詳細情報とコマンド構文については、`mkfs.gfs2` のマニュアルページを参照してください。

表 19.2: 重要なGFS2パラメータ

GFS2パラメータ	説明と推奨設定
ロック プロトコル 名 (<code>-p</code>)	使用するロックングプロトコルの名前。使用可能なロックングプロトコルは <code>lock_dlm</code> (共有ストレージ用) です。またはローカルファイルシステム(1 ノードのみ)としてGFS2を使用している場合は、 <code>lock_nolock</code> プロトコルを指定できます。このオプションを指定しない場合、 <code>lock_dlm</code> プロトコルであるとみなされます。
ロックテーブル名(<code>-t</code>)	使用しているロックモジュールに適切はロックテーブルフィールド。 <code>clustername: fsname</code> です。 <code>clustername</code> は、クラスタ設定ファイル <code>/etc/corosync/corosync.conf</code> のクラスタ名と一致している必要があります。このクラスタのメンバーだけが、このファイルシステムの使用を許可されます。 <code>fsname</code> は、このGFS2ファイルシステムと作成された他のファイルシステムを区別するために使用される固有のファイルシステム名です (1~16文字)。
ジャーナル数(<code>-j</code>)	作成する <code>gfs2_mkfs</code> 用のジャーナル数。ファイルシステムをマウントするマシンごとに少なくとも1つのジャーナルが必要です。このオプションを指定しない場合、1つのジャーナルが作成されます。

手順 19.2: GFS2ボリュームの作成とフォーマット

クラスタノードの1つだけで、次の手順を実行します。

1. 端末ウィンドウを開いて、`root` としてログインします。

2. クラスタがオンラインであることをコマンド `crm status` で確認します。
3. `mkfs.gfs2` ユーティリティを使用して、ボリュームを作成およびフォーマットします。このコマンドの構文については、`mkfs.gfs2` マニュアルページを参照してください。
たとえば、最大32台のクラスタノードをサポートする新しいGFS2ファイルシステムを `/dev/sdb1` 上に作成するには、次のコマンドを使用します。

```
root # mkfs.gfs2 -t hacluster:mygfs2 -p lock_dlm -j 32 /dev/sdb1
```

`hacluster` 名は、ファイル `/etc/corosync/corosync.conf` (これはデフォルトです)のエントリ `cluster_name` に関係します。

19.4 GFS2ボリュームのマウント

GFS2ボリュームは、手動でマウントするか、クラスタマネージャでマウントできます(手順19.4「クラスタマネージャによるGFS2ボリュームのマウント」を参照)。

手順 19.3: GFS2ボリュームの手動によるマウント

1. 端末ウィンドウを開いて、`root` としてログインします。
2. クラスタがオンラインであることをコマンド `crm status` で確認します。
3. コマンドラインから、`mount` コマンドを使ってボリュームをマウントします。



警告: 手動マウントによるGFS2デバイス

GFS2ファイルシステムをテスト目的で手動マウントした場合、そのファイルシステムは、いったんマウント解除してから、クラスタリソースで使用してください。

手順 19.4: クラスタマネージャによるGFS2ボリュームのマウント

High AvailabilityソフトウェアでGFS2ボリュームをマウントするには、クラスタ内でOCFファイルシステムのリソースを設定します。次の手順では、`crm` シェルを使用してクラスタリソースを設定します。リソースの設定には、Hawk2を使用することもできます。

1. シェルを起動し、`root` または同等のものとしてログインします。
2. `crm configure` を実行します。
3. GFS2ファイルシステムをクラスタ内のすべてのノードにマウントするように、Pacemakerを設定します。

```
crm(live)configure# primitive gfs2-1 ocf:heartbeat:Filesystem \  
  params device="/dev/sdb1" directory="/mnt/shared" fstype="gfs2" \  
  op monitor interval="20" timeout="40" \  
  op start timeout="60" op stop timeout="60" \  
  meta target-role="Stopped"
```

4. 手順17.1「DLMのベースグループの設定」で作成した dlm プリミティブと gfs2-1 プリミティブから構成されるベースグループを作成します。グループをクローンします。

```
crm(live)configure# group g-storage dlm gfs2-1  
  clone cl-storage g-storage \  
  meta interleave="true"
```

ベースグループの内部コロケーションおよび順序付けによって、Pacemakerは、すでに実行している dlm リソースも持つノード上で gfs2-1 リソースのみ起動します。

5. show で変更内容をレビューします。
6. すべて正しいければ、commit で変更を送信し、exit でcrmライブ設定を終了します。

20 DRBD

分散複製ブロックデバイス(DRBD*)を使用すると、IPネットワーク内の2つの異なるサイトに位置する2つのブロックデバイスのミラーを作成できます。Corosyncと共に使用すると、DRBDは分散高可用性Linuxクラスタをサポートします。この章では、DRBDのインストールとセットアップの方法を示します。

20.1 概念の概要

DRBDは、プライマリデバイス上のデータをセカンダリデバイスに、データの両方のコピーが同一に保たれるような方法で複製します。これは、ネットワーク型のRAID 1と考えてください。DRBDは、データをリアルタイムでミラーリングするので、そのレプリケーションは連続的に起こります。アプリケーションは、実際そのデータがさまざまなディスクに保存されるということを知る必要はありません。

DRBDは、Linuxカーネルモジュールであり、下端のI/Oスケジューラと上端のファイルシステムの間に存在しています(図20.1「Linux内でのDRBDの位置」参照)。DRBDと通信するには、高レベルのコマンド `drbdadm` を使用します。柔軟性を最大にするため、DRBDには、低レベルのツール `drbdsetup` が付いてきます。

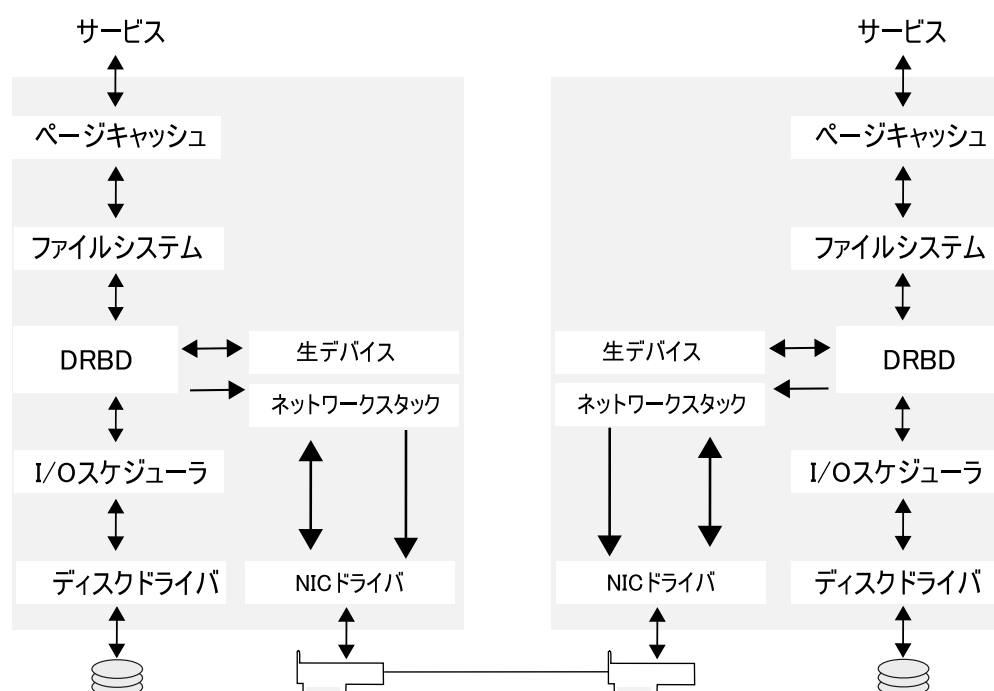


図 20.1: LINUX内でのDRBDの位置

！ 重要: 暗号化されないデータ

ミラー間のデータトラフィックは暗号化されません。データ交換を安全にするには、接続に仮想プライベートネットワーク(VPN)ソリューションを導入する必要があります。

DRBDでは、Linuxでサポートされる任意のブロックデバイスを使用できます。通常は次のデバイスです。

- パーティションまたは完全なハードディスク
- ソフトウェアRAID
- LVM (Logical Volume Manager)
- EVMS (Enterprise Volume Management System)

DRBDは、デフォルトでは、DRBDノード間の通信にTCPポート 7788 以上を使用します。使用しているポートの通信がファイアウォールで許可されていることを確認してください。

まず、DRBDデバイスを設定してから、その上にファイルシステムを作成する必要があります。ユーザデータに関することはすべて、rawデバイスではなく、/dev/drbdN デバイスを介してのみ実行される必要があります。これは、DRBDが、メタデータ用にrawデバイスの最後の部分を使用するからです。rawデバイスを使用すると、データが矛盾する原因となります。

udevの統合により、/dev/drbd/by-res/RESOURCES の形式でシンボリックリンクも取得されます。このリンクは、より簡単に使用でき、デバイスのマイナー番号を誤って記憶しないように安全対策が講じられています。

たとえば、rawデバイスのサイズが1024MBの場合、DRBDデバイスは、1023MBしかデータ用に使用できません。70KBは隠され、メタデータ用に予約されています。rawディスクを介した既存のキロバイトへのアクセスは、それがユーザデータ用でないので、すべて失敗します。

20.2 DRBDサービスのインストール

パートI「インストール、セットアップ、およびアップグレード」で説明されているように、High Availability Extensionをネットワーククラスターの両方のSUSE Linux Enterprise Serverマシンにインストールします。High Availability Extensionをインストールすると、DRBDプログラムファイルもインストールされます。

クラスタスタック全体を必要とせず、のみを使用したい場合、パッケージ drbd、drbd-kmp-FLAVOR、drbd-utils、および yast2-drbd をインストールしてください。

drbdadm の操作を簡素化するには、Bash補完サポートを使用します。現在のシェルセッションでこのサポートを有効にするには、次のコマンドを挿入します。

```
root # source /etc/bash_completion.d/drbdadm.sh
```

root 用に永続的に使用するには、ファイル /root/.bashrc を拡張し、前の行を挿入します。

20.3 DRBDサービスの設定



注記: 必要な調整

次の手順では、サーバ名としてaliceとbobを使用し、クラスタリソース名として r0 を使用します。aliceをプライマリノードとして設定し、/dev/sda1 をストレージとして設定します。必ず、手順を変更して、ご使用のノード名とファイルの名前を使用してください。

次の項では、aliceとbobという2つのノードがあり、それぞれがTCPポート 7788 を使用するものと想定しています。ファイアウォールでこのポートが開いているようにしてください。

1. システムを準備します。

- a. Linuxノード内のブロックデバイスを準備し、(必要な場合は)パーティション分割しておいてください。
- b. ディスクに、必要のなくなったファイルシステムがすでに含まれている場合は、次のコマンドでファイルシステムの構造を破壊します。

```
root # dd if=/dev/zero of=YOUR_DEVICE count=16 bs=1M
```

破壊する、より多くのファイルシステムがある場合は、DRBDセットアップに含むすべてのデバイス上でこのステップを繰り返します。

- c. クラスタがすでにDRBDを使用している場合は、クラスタを保守モードにします。

```
root # crm configure property maintenance-mode=true
```

クラスタがすでにDRBDを使用している場合に、この手順をスキップすると、ライブ設定の構文エラーによってサービスがシャットダウンされます。

別の方法として、drbdadm -c FILE を使用して設定ファイルをテストすることもできます。

2. 次のいずれかの方法を選択してDRBDを設定します。

- 20.3.1項「手動によるDRBDの設定」
- 20.3.2項「YaSTによるDRBDの設定」

3. Csync2 (デフォルト)を設定している場合、DRBD設定ファイルは、同期に必要なファイルのリストにすでに含まれています。同期するには、次のコマンドを使用します。

```
root # csync2 -xv /etc/drbd.d/
```

Csync2を設定していない場合(または使用しない場合)には、DRBD設定ファイルを手動で他のノードにコピーしてください。

```
root # scp /etc/drbd.conf bob:/etc/
root # scp /etc/drbd.d/* bob:/etc/drbd.d/
```

4. 初期同期を実行します(20.3.3項「DRBDリソースの初期化とフォーマット」を参照してください)。
5. クラスタの保守モードフラグをリセットします。

```
root # crm configure property maintenance-mode=false
```

20.3.1 手動によるDRBDの設定



注記: 「自動プロモート」機能の限定サポート

DRBD9機能の「自動プロモート」は、マスタ/スレーブ接続の代わりにクローンとファイルシステムリソースを使用できます。ファイルシステムがマウントされているときにこの機能を使用すると、DRBDは自動的にプライマリモードに変わります。

自動プロモート機能は現在サポートが限定されています。DRBD 9では、SUSEはDRBD-8でもサポートされていた使用例と同じ使用例をサポートしています。3つ以上のノードでのセットアップなど、それを超える使用例はサポートされていません。

DRBDを手動で設定するには、次の手順に従います。

手順 20.1: DRBDの手動設定

DRBDバージョン8.3以降、設定ファイルは、複数のファイルに分割され、/etc/drbd.d/ ディレクトリに保存されています。

1. /etc/drbd.d/global_common.conf ファイルを開きます。このファイルには、すでにいくつかのグローバルな事前定義値が含まれています。startup セクションに移動し、次の3行を挿入します。

```
startup {
```



```
# wfc-timeout degr-wfc-timeout outdated-wfc-timeout
# wait-after-sb;
wfc-timeout 100;
degr-wfc-timeout 120;
}
```

これらのオプションは、ブート時のタイムアウトを減らすために使用します。詳細については、<https://docs.linbit.com/docs/users-guide-9.0/#ch-configure> を参照してください。

2. ファイル `/etc/drbd.d/r0.res` を作成し、状況に合わせて行を変更し、ファイルを保存します。

```
resource r0 { ❶
    device /dev/drbd0; ❷
    disk /dev/sda1; ❸
    meta-disk internal; ❹
    on alice { ❺
        address 192.168.1.10:7788; ❻
        node-id 0; ❼
    }
    on bob { ❺
        address 192.168.1.11:7788; ❻
        node-id 1; ❼
    }
    disk {
        resync-rate 10M; ❸
    }
    connection-mesh { ❹
        hosts alice bob;
    }
}
```

- ❶ 必要なサービスへのいくつかの関連付けを許可するDRBDリソースの名前。たとえば、`nfs`、`http`、`mysql_0`、`postgres_wal` など。
- ❷ DRBD用デバイス名とそのマイナー番号。
先に示した例では、マイナー番号0がDRBDに対して使用されています。udev統合スクリプトは、シンボリックリンク(`/dev/drbd/by-res/nfs/0`)を提供します。または、設定のデバイスノード名を省略し、代わりに次のラインを使用します。
`drbd0 minor 0` (`/dev/` は必要に応じて指定します)または `/dev/drbd0`
- ❸ ノード間で複製されるrawデバイス。ただし、この例では、デバイスは両方のノードで「同じ」です。異なるデバイスが必要な場合は、`disk` パラメータを `on` ホストに移動します。
- ❹ `meta-disk` パラメータには、通常、値 `internal` が含まれますが、メタデータを保持する明示的なデバイスを指定することもできます。詳細については、<https://docs.linbit.com/docs/users-guide-9.0/#s-metadata> を参照してください。
- ❺ `on` セクションでは、この設定文が適用されるホストを記述します。

- ⑥ それぞれのノードのIPアドレスとポート番号。リソースごとに、通常、7788 から始まる別個のポートが必要です。1つのDRBDリソースに対して両方のポートが同じである必要があります。
 - ⑦ 複数のノードを設定する際は、ノードIDが必要です。ノードIDは、別々のノードを区別するための固有の負でない整数です。
 - ⑧ 同期レート。このレートは、ディスク帯域幅およびネットワーク帯域幅の3分の1に設定します。これは、再同期を制限するだけで、レプリケーションは制限しません。
 - ⑨ 同一メッシュのすべてのノードを設定します。hosts パラメータには、同じDRBDセットアップを共有するすべてのホスト名が含まれます。
3. 環境設定ファイルの構文をチェックします。次のコマンドがエラーを返す場合は、ファイルを検証します。
- ```
root # drbdadm dump all
```
4. [20.3.3項「DRBDリソースの初期化とフォーマット」](#)に進みます。

## 20.3.2 YaSTによるDRBDの設定

YaSTを使用して、DRBDの初期セットアップを開始できます。DRBDセットアップの作成後、生成されたファイルを手動で調整できます。

ただし、設定ファイルを変更した後にYaST DRBDモジュールを使用しないでください。DRBDモジュールでサポートされているのは、限られた一連の基本設定のみです。再度DRBDモジュールを使用すると、変更内容がモジュールに表示されない可能性があります。

YaSTを使ってDRBDを設定するには、次の手順に従います。

### 手順 20.2: YASTを使用してDRBDを設定

1. YaSTを起動して、設定モジュール[高可用性] > [DRBD]を選択します。すでにDRBDを設定していた場合、YaSTはそのことを警告します。YaSTは設定を変更し、古いDRBD設定ファイルを \*.YaSTsave として保存します。
2. [起動設定] > [ブート]のブートフラグは、そのままにしてください(デフォルトでは off)。Pacemakerがこのサービスを管理するので変更しないでください。
3. ファイアウォールが実行中の場合は、[ファイアウォールでポートを開く]を有効にします。
4. [リソース設定]エントリに移動します。[追加]をクリックして、新しいリソースを作成します([図 20.2「リソースの環境設定」](#)を参照してください)。

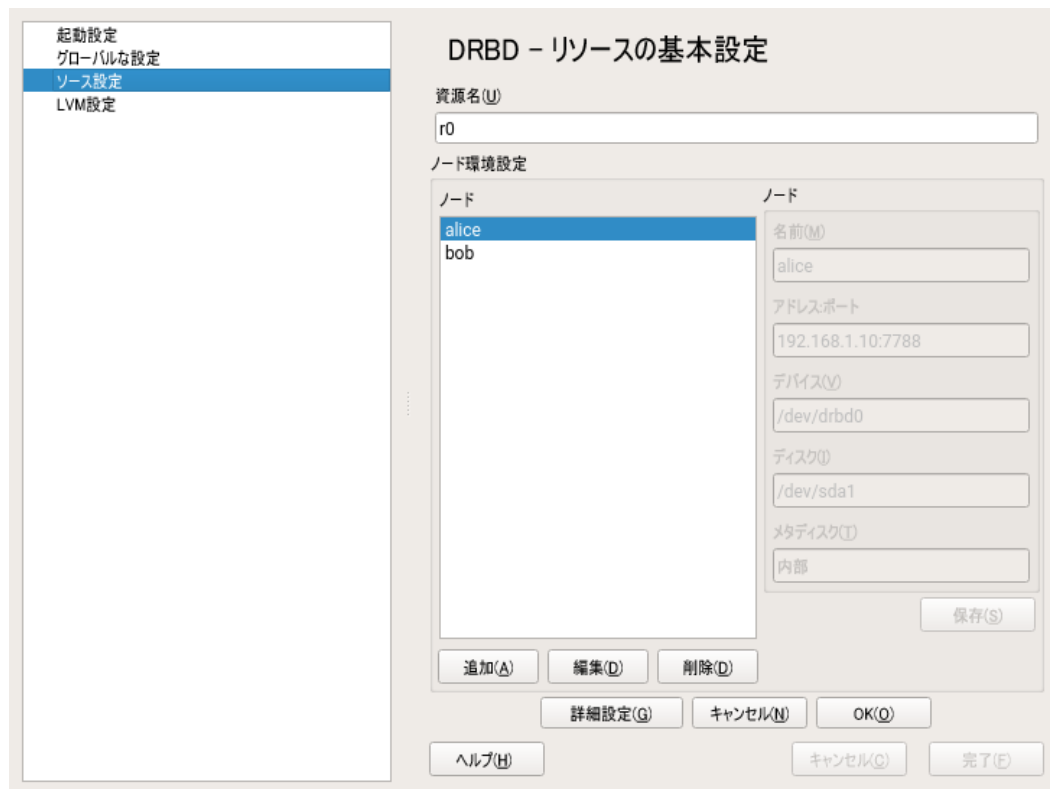


図 20.2: リソースの環境設定

次のパラメータを設定する必要があります。

[リソース名]

DRBDリソースの名前(必須)。

[Name (名前)]

関連するノードのホスト名。

[アドレス:ポート]

それぞれのノードのIPアドレスとポート番号(デフォルトは 7788)

[デバイス]

複製されたデータにアクセスするためのブロックデバイスパス。デバイスにマイナー番号が使用されている場合は、関連付けられたブロックデバイスの名前は /dev/drbdX になることが普通です。Xはデバイスのマイナー番号です。デバイスにマイナー番号が使用されていない場合は、必ずデバイス名の後に minor 0を追記します。

#### [ディスク]

両方のノード間で複製されるrawデバイス。LVMを使用する場合、LVMデバイス名を挿入します。

#### [メタディスク]

[メタディスク]は、値 `internal` に設定されるか、またはインデックスで拡張された、drbdで必要なメタデータを保持する明示的なデバイスを指定します。

複数のdrbdリソースに実際のデバイスを使用することもできます。たとえば、最初のリソースに対して[メタディスク]が `/dev/sda6[0]` の場合、`/dev/sda6[1]` を2番目のリソースに使用できます。ただし、このディスク上で各リソースについて少なくとも128MBのスペースが必要です。メタデータの固定サイズによって、複製できる最大データサイズが制限されます。

これらのオプションはすべて、`/usr/share/doc/packages/drbd/drbd.conf` ファイルの例と `drbd.conf(5)` のマニュアルページで説明されています。

5. [保存]をクリックします。
6. [追加]をクリックして、2番目のDRBDリソースを入力し、[保存]をクリックして終了します。
7. [OK]と[完了]をクリックして、[リソース設定]を閉じます。
8. DRBDでLVMを使用する場合、LVM設定ファイルでオプションをいくつか変更する必要があります([LVM Configuration (LVMの設定)]エントリを参照してください)。この変更は、YaST DRBDモジュールを使用して自動的に実行できます。  
DRBDリソースのローカルホストのディスク名およびデフォルトのフィルタはLVMフィルタで拒否されます。LVMデバイスをスキャンできるのは `/dev/drbd` のみです。  
たとえば、`/dev/sda1` をDRBDディスクとして使用している場合、そのデバイス名がLVMフィルタの最初のエントリとして挿入されます。フィルタを手動で変更するには、[Modify LVM Device Filter Automatically (LVMデバイスフィルタを自動的に変更)]チェックボックスをクリックします。
9. [完了]をクリックして、変更を保存します。
10. 20.3.3項「DRBDリソースの初期化とフォーマット」に進みます。

### 20.3.3 DRBDリソースの初期化とフォーマット

システムを準備してDRBDを設定したら、ディスクの初回の初期化を行います。

1. 両ノード(aliceとbob)でメタデータストレージを初期化します。

```
root # drbdadm create-md r0
root # drbdadm up r0
```

2. DRBDリソースの初期再同期を短縮する場合は、次のことを確認します。

- すべてのノード上のDRBDデバイスが同じデータを持つ場合(たとえば、20.3項「DRBD サービスの設定」に示すように `dd` コマンドでファイルシステム構造を破壊することによる)、次のコマンドを使用して初期再同期をスキップします(両ノード)。

```
root # drbdadm new-current-uuid --clear-bitmap r0/0
```

状態は Secondary/Secondary UpToDate/UpToDate です

- その後、次のステップに進みます。

3. プライマリノードのaliceから再同期プロセスを開始します。

```
root # drbdadm primary --force r0
```

4. 以下を使用してステータスをチェックします。

```
root # drbdadm status r0
r0 role:Primary
 disk:UpToDate
 bob role:Secondary
 peer-disk:UpToDate
```

5. DRBDデバイスの上にファイルシステムを作成します。たとえば、次のように指定します。

```
root # mkfs.ext3 /dev/drbd0
```

6. ファイルシステムをマウントして使用します。

```
root # mount /dev/drbd0 /mnt/
```

## 20.4 DRBD 8から DRBD 9への移行

DRBD 8 (SUSE Linux Enterprise High Availability Extension 12 SP1に付属)とDRBD 9 (SUSE Linux Enterprise High Availability Extension 12 SP2に付属)間で、メタデータフォーマットが変更されました。DRBD 9では、以前のメタデータファイルが新しいフォーマットに自動的に変換されません。

12 SP2に移行した後で、DRBDを開始する前に、DRBDメタデータをバージョン9フォーマットに手動で変換します。これを実行するには、`drbdadm create-md`を使用します。設定を変更する必要はありません。



## 注記: 限定サポート

DRBD 9では、SUSEはDRBD-8でもサポートされていた使用例と同じ使用例をサポートしています。3つ以上のノードでのセットアップなど、それを超える使用例はサポートされていません。

DRBD 9は、バージョン8と互換性を持つように切り替わります。3つ以上のノードの場合、DRBDバージョン9固有のオプションを使用するため、メタデータを再作成する必要があります。

スタックされたDRBDリソースがある場合は、詳細について[20.5項「スタックされたDRBDデバイスの作成」](#)も参照してください。

新しいリソースを再作成せずにデータを保持し、新しいノードを追加することを許可する場合は、次の操作を実行します。

1. 1つのノードをスタンバイモードで設定します。
2. ノードのすべてでDRBDパッケージのすべてを更新します。[20.2項「DRBDサービスのインストール」](#)を参照してください。
3. リソース設定に新しいノード情報を追加します。
  - `すべての on` セクションごとにnode-id。
  - `connection-mesh` セクションには、ホストパラメータのすべてのホスト名が含まれます。`hosts` パラメータで制御します。

[手順20.1「DRBDの手動設定」](#)のサンプル設定を参照してください。

4. `internal` を `meta-disk` キーとして使用する場合は、DRBDディスクのスペースを拡大します。LVMのようなスペースの拡大をサポートするデバイスを使用します。別の方法として、メタデータ用の外部ディスクに変更して、`meta-disk DEVICE`; を使用します。
5. 新しい設定に基づいてメタデータを再作成します。

```
root # drbdadm create-md RESOURCE
```

6. スタンバイモードをキャンセルします。

## 20.5 スタックされたDRBDデバイスの作成

スタックされたDRBDデバイスには少なくとも一方のデバイスがDRBDリソースでもある2つの他のデバイスが含まれます。すなわち、DRBDは既存のDRBDリソースの最上部に追加のノードを追加します(図20.3「リソースのスタッキング」を参照してください)。このようなレプリケーションセットアップは、バックアップおよび障害復旧目的に使用できます。

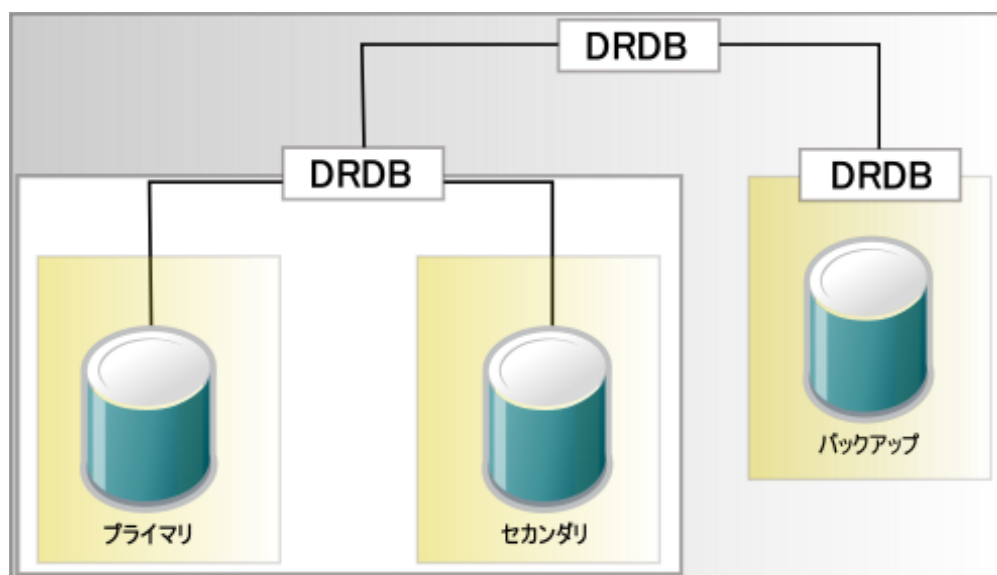


図 20.3: リソースのスタッキング

Three-wayレプリケーションは、非同期(DRBDプロトコルA)と同期レプリケーション(DRBDプロトコルC)を使用します。非同期部分はスタックされたリソースに使用され、同期部分はバックアップに使用されます。

ご使用の運用環境ではスタックされたデバイスを使用します。たとえば、最上部にDRBDデバイス `/dev/drbd0` とスタックされたデバイス `/dev/drbd10` がある場合、ファイルシステムは `/dev/drbd10` 上に作成されます。詳細については、例20.1「3ノードのスタックされたDRBDリソースの設定」を参照してください。

例 20.1: 3ノードのスタックされたDRBDリソースの設定

```
/etc/drbd.d/r0.res
resource r0 {
 protocol C;
 device /dev/drbd0;
 disk /dev/sda1;
 meta-disk internal;

 on amsterdam-alice {
 address 192.168.1.1:7900;
 }
}
```

```

on amsterdam-bob {
 address 192.168.1.2:7900;
}

resource r0-U {
 protocol A;
 device /dev/drbd10;

 stacked-on-top-of r0 {
 address 192.168.2.1:7910;
 }

 on berlin-charlie {
 disk /dev/sda10;
 address 192.168.2.2:7910; # Public IP of the backup node
 meta-disk internal;
 }
}

```

## 20.6 リソースレベルのフェンシングの使用

DRBDレプリケーションリンクが途切れた場合、PacemakerはDRBDリソースを別のノードにプロモートしようとしています。Pacemakerが古いデータでサービスを開始しないようにするため、[例20.2「クラスタ情報ベース\(CIB\)を使用したリソースレベルのフェンシングを含むDRBDの設定」](#)に示すようにDRBD設定ファイルでリソースレベルのフェンシングを有効にします。

例 20.2: クラスタ情報ベース(CIB)を使用したリソースレベルのフェンシングを含むDRBDの設定

```

resource RESOURCE {
 net {
 fencing resource-only;
 # ...
 }
 handlers {
 fence-peer "/usr/lib/drbd/crm-fence-peer.9.sh";
 after-resync-target "/usr/lib/drbd/crm-unfence-peer.9.sh";
 # ...
 }
 ...
}

```

DRBDレプリケーションリンクが切断されると、DRBDは以下を実行します。

1. DRBDは `crm-fence-peer.9.sh` スクリプトを呼び出します。
2. スクリプトはクラスタマネージャに連絡します。



3. スクリプトはこのDRBDリソースに関連付けられたPacemakerリソースを判断します。
4. スクリプトは、このDRBDリソースが他のノードにプロモートされないことを確認します。リソースは現在アクティブなノード上にとどまります。
5. レプリケーションリンクがもう一度接続され、DRBDがその同期プロセスを完了すると、制限が解除されます。クラスタマネージャは自由にリソースをプロモートできるようになります。

## 20.7 DRBDサービスのテスト

インストールと設定のプロシージャが予期どおりの結果となった場合は、DRBD機能の基本的なテストを実行できます。このテストは、DRBDソフトウェアの機能を理解する上でも役立ちます。

1. alice上でDRBDサービスをテストします。

- a. 端末コンソールを開き、rootとしてログインします。
- b. aliceにマウントポイント(/srv/r0など)を作成します。

```
root # mkdir -p /srv/r0
```

- c. drbd デバイスをマウントします。

```
root # mount -o rw /dev/drbd0 /srv/r0
```

- d. プライマリノードからファイルを作成します。

```
root # touch /srv/r0/from_alice
```

- e. aliceでディスクをマウント解除します。

```
root # umount /srv/r0
```

- f. aliceで次のコマンドを入力して、aliceのDRBDサービスを降格します。

```
root # drbdadm secondary r0
```

2. bob上でDRBDサービスをテストします。

- a. 端末コンソールを開き、bobで rootとしてログインします。
- b. bobで、DRBDサービスをプライマリに昇格します。

```
root # drbdadm primary r0
```

- c. bobで、bobがプライマリかどうかチェックします。

```
root # drbdadm status r0
```

- d. bobで、/srv/r0などのマウントポイントを作成します。

```
root # mkdir /srv/r0
```

- e. bobで、DRBDデバイスをマウントします。

```
root # mount -o rw /dev/drbd0 /srv/r0
```

- f. aliceで作成したファイルが存在していることを確認します。

```
root # ls /srv/r0/from_alice
```

/srv/r0/from\_alice ファイルが一覧に表示されている必要があります。

3. サービスが両方のノードで稼動していれば、DRBDの設定は完了です。

4. 再度、aliceをプライマリとして設定します。

- a. bobで次のコマンドを入力して、bobのディスクをマウント解除します。

```
root # umount /srv/r0
```

- b. bobで次のコマンドを入力して、bobのDRBDサービスを降格します。

```
root # drbdadm secondary
```

- c. aliceで、DRBDサービスをプライマリに昇格します。

```
root # drbdadm primary
```


- d. aliceで、aliceがプライマリかどうかチェックします。

```
root # drbdadm status r0
```

5. サービスを自動的に起動させ、サーバに問題が発生した場合はフェールオーバーさせるためには、Pacemaker/CorosyncでDRBDを高可用性サービスとして設定できます。SUSE Linux Enterprise 12 SP5のインストールと設定については、[パートII「設定および管理」](#)を参照してください。

## 20.8 DRBDのチューニング

DRBDをチューニングするには、いくつかの方法があります。

1. メタデータ用には外部ディスクを使用します。これは便利ですが、保守作業は煩雑になります。
2. `sysctl`を介して受信および送信バッファ設定を変更することで、ネットワーク接続を調整します。
3. DRBD設定で `max-buffers`、`max-epoch-size`、またはその両方を変更します。
4. IOパターンに応じて、`al-extents` の値を増やします。
5. ハードウェアRAIDコントローラとBBU (「バッテリーバックアップユニット」)を併用する場合、`no-disk-flushes`、`no-disk-barrier`、および `no-md-flushes` の設定が有効な場合があります。
6. ワークロードに従って読み込みバランスを有効にします。詳細については、<https://www.linbit.com/en/read-balancing/>  を参照してください。

## 20.9 DRBDのトラブルシューティング

DRBDセットアップには、多数のコンポーネントが使用され、別のソースから問題が発生することがあります。以降のセクションでは、一般的なシナリオをいくつか示し、さまざまなソリューションを推奨します。

### 20.9.1 環境設定

初期のDRBDセットアップが予期どおりに機能しない場合は、おそらく、環境設定に問題があります。環境設定の情報を取得するには:

1. 端末コンソールを開き、`root`としてログインします。
2. `drbdadm`に `-d` オプションを指定して、環境設定ファイルをテストします。入力次のコマンドを入力します。

```
root # drbdadm -d adjust r0
```

`adjust` オプションのドライ実行では、`drbdadm`は、DRBDリソースの実際の設定を使用中のDRBD環境設定ファイルと比較しますが、コールは実行しません。出力をレビューして、エラーのソースおよび原因を確認してください。

3. `/etc/drbd.d/*` ファイルと `drbd.conf` ファイルにエラーがある場合は、そのエラーを修正してから続行してください。
4. パーティションと設定が正しい場合は、`drbdadm` を `-d` オプションなしで、再度実行します。

```
root # drbdadm adjust r0
```

このコマンドは、環境設定ファイルをDRBDリソースに適用します。

## 20.9.2 ホスト名

DRBDの場合、ホスト名の太文字と小文字が区別され(`Node0` は `node0` とは異なるホストであるとみなされる)、カーネルに格納されているホスト名と比較されます(`uname -n` 出力を参照)。

複数のネットワークデバイスがあり、専用ネットワークデバイスを使用したい場合、おそらく、ホスト名は使用されたIPアドレスに解決されません。この場合は、パラメータ `disable-ip-verification` を使用します。

## 20.9.3 TCPポート7788

システムがピアに接続できない場合は、ローカルファイアウォールに問題のある可能性があります。DRBDは、デフォルトでは、TCPポート `7788` を使用して、もう一方のノードにアクセスします。このポートを両方のノードからアクセスできるかどうか確認してください。

## 20.9.4 DRBDデバイスが再起動後に破損した

DRBDサブシステムが実際のどのデバイスが最新データを保持しているか認識していない場合、スプリットブレイン受験に変更されます。この場合、それぞれのDRBDサブシステムがセカンダリとして起動され、互いに接続しません。この場合、ログ記録データに、次のメッセージが出力されることがあります。

```
Split-Brain detected, dropping connection!
```

この状況を解決するには、廃棄するデータを持つノードで、次のコマンドを入力します。

```
root # drbdadm secondary r0
```

状態が `WFconnection` の場合、最初に切断します。

```
root # drbdadm disconnect r0
```

最新のデータを持つノードで、次のコマンドを入力します。

```
root # drbdadm connect --discard-my-data r0
```

このコマンドは、あるノードのデータをピアのデータで上書きすることによって問題を解決するため、両方のノードで一貫したビューが得られます。

## 20.10 詳細情報

DRBDについては、次のオープンソースリソースを利用できます。

- プロジェクトホームページ<http://www.drbd.org>。
- 詳細については、項目「Highly Available NFS Storage with DRBD and Pacemaker」を参照してください。
- [http://clusterlabs.org/wiki/DRBD\\_HowTo\\_1.0](http://clusterlabs.org/wiki/DRBD_HowTo_1.0) (Linux Pacemaker Cluster Stack Projectによる)。
- このディストリビューションで利用できるDRBDのマニュアルページは、`drbd(8)`、`rbdmeta(8)`、`drbdsetup(8)`、`drbdadm(8)`、`drbd.conf(5)` です。
- コメント付きのDRBD設定例が、`/usr/share/doc/packages/drbd-utils/drbd.conf.example` にあります。
- さらに、クラスタ間のストレージ管理を容易にするために、「DRBD-Manager」(<https://www.linbit.com/en/drbd-manager/>)に関する最新の通知を参照してください。

## 21 Cluster Logical Volume Manager(cLVM)

クラスタ上の共有ストレージを管理する場合、ストレージサブシステムに行った変更を各ノードに伝える必要があります。Logical Volume Manager 2 (LVM2)はローカルストレージの管理に多用されており、クラスタ全体のボリュームグループのトランスペアレントな管理をサポートするために拡張されています。クラスタ化されたボリュームグループを、ローカルストレージと同じコマンドで管理できます。

### 21.1 概念の概要

クラスタLVM2は、さまざまなツールと連携します。

分散ロックマネージャ(DLM:Distributed Lock Manager)

ロックを通じてcLVMのディスクアクセスとメタデータへのアクセスを調整します。

論理ボリュームマネージャ2(LVM2: Logical Volume Manager2)

1つのファイルシステムをいくつかのディスクに柔軟に分散することができます。LVM2は、ディスクスペースの仮想プールを提供します。

クラスタ化論理ボリュームマネージャ(cLVM: Clustered Logical Volume Manager)

すべてのノードが変更を知ることができるように、LVMメタデータへのアクセスを調整します。cLVMは、共有データ自体へのアクセスは調整しません。これをcLVMができるようにするには、OCFS2などのクラスタ対応アプリケーションをcLVMの管理対象ストレージの上に設定する必要があります。

### 21.2 cLVMの環境設定

ご使用のシナリオによっては、次のレイヤを使用して、cLVMでRAID 1デバイスを作成することができます。

- **LVM2:** ファイルシステムのサイズを増減したり、物理ストレージを追加したり、ファイルシステムのスナップショットを作成する場合に、高い柔軟性を提供するソリューションです。この方法については、[21.2.3項「シナリオ - SAN上でiSCSIを使用するcLVM」](#)に説明があります。
- **DRBD:** RAID 0 (ストライピング)とRAID 1 (ミラーリング)のみを提供します。最後の方式については、[21.2.4項「シナリオ - DRBDを使用するcLVM」](#)に説明があります。

次の前提条件を満たしていることを確認してください。

- 共有ストレージデバイス(Fibre Channel、FCoE、SCSI、iSCSI SAN、DRBD\*で提供されているデバイスなど)が使用できること
- DRBDの場合は、両方のノードがプライマリであること(以降の手順で説明)。
- LVM2のロックタイプがクラスタを認識するかどうか確認すること。`/etc/lvm/lvm.conf` 内のキーワード `locking_type` に値 `3` が含まれている必要があります(デフォルトは `1` です)。必要な場合は、この設定をすべてのノード にコピーします。
- cLVMで使用できないため、`lvm` デーモンが無効になっているかどうかを確認します。`/etc/lvm/lvm.conf` で、キーワード `use_lvm` が `0` に設定される必要があります(デフォルトは `1` です)。必要な場合は、この設定をすべてのノード にコピーします。

## 21.2.1 クラスタリソースの作成

cLVMを使用するためのクラスタ準備には次の基本的な手順が含まれます。

- DLMリソースを作成する
- DLM、CLVM、およびSTONITHの設定

手順 21.1: DLMリソースを作成する

1. シェルを起動して、`root` としてログインします。
2. クラスタリソースの現在の設定を確認します。

```
root # crm configure show
```

3. すでにDLMリソース(および対応するベースグループおよびベースクローン)を設定済みである場合、[手順21.2「DLM、CLVM、およびSTONITHの設定」](#)で続きます。  
そうでない場合は、[手順17.1「DLMのベースグループの設定」](#)で説明されているように、DLMリソース、および対応するベースグループとベースクローンを設定します。
4. `crm` ライブ設定を `exit` で終了します。

## 21.2.2 シナリオ: Cmirrordの設定

クラスタのミラーログ情報を追跡するには、`cmirrord` デーモンを使用します。このデーモンが実行されていないと、クラスタはミラーリングできません。

`/dev/sda` と `/dev/sdb` は、DRBD、iSCSI、その他と同様の共有ストレージデバイスであると想定します。必要な場合は、これらを独自のデバイス名に置き換えます。次の手順に従います。

## 手順 21.2: DLM、CLVM、およびSTONITHの設定

1. 2つ以上のノードを持つクラスタの作成方法については、『インストールおよびセットアップガイド クラスタ』を参照してください。
2. `dlm`、`clvmd`、およびSTONITHを実行するために、クラスタを構成します。

```
root # crm configure
crm(live)configure# primitive clvmd ocf:heartbeat:clvm \
 params with_cmirrord=1 \
 op stop interval=0 timeout=100 \
 op start interval=0 timeout=90 \
 op monitor interval=20 timeout=20
crm(live)configure# primitive dlm ocf:pacemaker:controld \
 op start timeout="90" \
 op stop timeout="100" \
 op monitor interval="60" timeout="60"
crm(live)configure# primitive sbd_stonith stonith:external/sbd \
 params pcmk_delay_max=30
crm(live)configure# group g-storage dlm clvmd
crm(live)configure# clone cl-storage g-storage \
 meta interleave="true" ordered=true
```

3. `exit` で`crmsh`を終了し、変更内容をコミットします。

手順 21.3を使用してディスクの構成を続行します。

## 手順 21.3: CLVM用のディスクの構成

1. クラスタ化されたボリュームグループ (VG) を作成します。

```
root # pvcreate /dev/sda /dev/sdb
root # vgcreate -cy vg1 /dev/sda /dev/sdb
```

2. ミラーログの論理ボリューム (LV) をクラスタ内に作成します。

```
root # lvcreate -n lv1 -m1 -l10%VG vg1 --mirrorlog mirrored
```

3. `lvs` を使用して進捗状況を表示します。パーセンテージの数値が100%に到達したら、ミラーディスクは正しく同期化されたということです。
4. クラスタ化されたボリューム `/dev/vg/lv1` をテストするには、次の手順に従います。
  - a. `/dev/vg/lv1` を読み込むか、ここに書き込みます。
  - b. `lvchange -an` でLVを非アクティブ化します。
  - c. `lvchange -ay` でLVをアクティブ化します。



d. `lvconvert` を使用してミラーログをディスクログに変換します。

5. 別のクラスタVGにミラーログのLVを作成します。これは前のものとは別のボリュームグループです。

現在のcLVMは、ミラーサイドごとに1つの物理ボリューム (PV) しか処理できません。1つのミラーが実際には、連結またはストライプ化の必要がある複数のPVで構成されている場合、`lvcreate` はこのことを理解できません。このため、`lvcreate` および `cmirrord` メタデータは、複数のPVを1つのサイドに「グループ化」することを理解する必要があり、事実上RAID10をサポートすることになります。

`cmirrord` に対してRAID10をサポートするには、次の手順を使用します (`/dev/sda`、`/dev/sdb`、`/dev/sdc`、および `/dev/sdd` は共有ストレージデバイスだとします)。

1. ボリュームグループ (VG) を作成します。

```
root # pvcreate /dev/sda /dev/sdb /dev/sdc /dev/sdd
Physical volume "/dev/sda" successfully created
Physical volume "/dev/sdb" successfully created
Physical volume "/dev/sdc" successfully created
Physical volume "/dev/sdd" successfully created
root # vgcreate vgtest /dev/sda /dev/sdb /dev/sdc /dev/sdd
Clustered volume group "vgtest" successfully created
```

2. ファイル `/etc/lvm/lvm.conf` を開き、`allocation` セクションに移動します。次の行を設定して、ファイルを保存します。

```
mirror_logs_require_separate_pvs = 1
```

3. PVにタグを追加します。

```
root # pvchange --addtag @a /dev/sda /dev/sdb
root # pvchange --addtag @b /dev/sdc /dev/sdd
```

タグは、ストレージオブジェクトのメタデータに割り当てられ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LVM2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になるように分類できます。

4. タグを一覧します。

```
root # pvs -o pv_name,vg_name,pv_tags /dev/sd{a,b,c,d}
```

次の出力を受信します。

| PV       | VG     | PV Tags |
|----------|--------|---------|
| /dev/sda | vgtest | a       |

```
/dev/sdb vgtest a
/dev/sdc vgtest b
/dev/sdd vgtest b
```

LVM2に関する詳細情報が必要な場合は、『SUSE Linux Enterprise Server 12 SP5 ストレージ管理ガイド』(<https://documentation.suse.com/sles-12/html/SLES-all/cha-lvm.html>)を参照してください。

### 21.2.3 シナリオ - SAN上でiSCSIを使用するcLVM

次のシナリオでは、iSCSIターゲットをいくつかのクライアントにエクスポートする2つのSANボックスを使用します。一般的なアイデアが、[図21.1「cLVMによるiSCSIのセットアップ」](#)で説明されています。

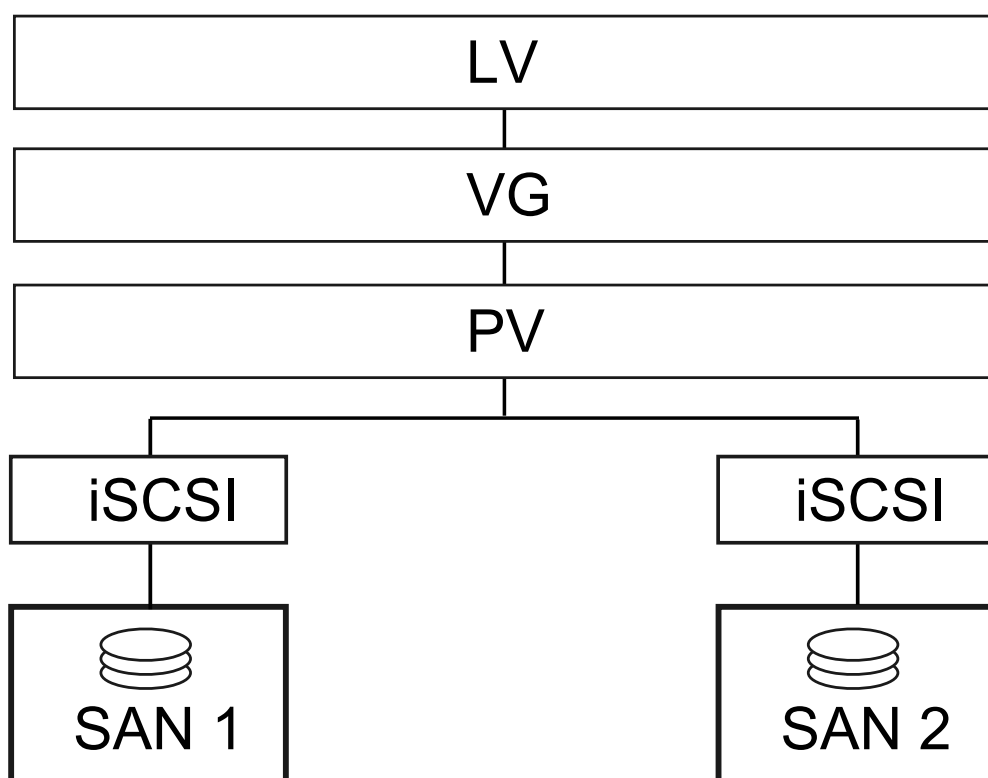


図 21.1: CLVMによるiSCSIのセットアップ



#### 警告: データ損失

以降の手順を実行すると、ディスク上のデータはすべて破壊されます。

まず、1つのSANボックスだけ設定します。各SANボックスは、そのiSCSIターゲットをエクスポートする必要があります。次の手順に従います。

#### 手順 21.4: iSCSIターゲット(SAN上)を設定する

1. YaSTを実行し、[ネットワークサービス] > [iSCSI LIO Target (iSCSI LIOターゲット)]の順にクリックしてiSCSIサーバモジュールを起動します。
2. コンピュータがブートするたびにiSCSIターゲットを起動したい場合は、[ブート時]を選択し、そうでない場合は、[手動]を選択します。
3. ファイアウォールが実行中の場合は、[ファイアウォールでポートを開く]を有効にします。
4. [グローバル]タブに切り替えます。認証が必要な場合は、受信または送信(あるいはその両方)の認証を有効にします。この例では、[認証なし]を選択します。
5. 新しいiSCSIターゲットを追加します。
  - a. [ターゲット]タブに切り替えます。
  - b. [追加]をクリックします。
  - c. ターゲットの名前を入力します。名前は、次のようにフォーマットされます。

```
iqn.DATE.DOMAIN
```

フォーマットに関する詳細は、セクション3.2.6.3.1のタイプ「iqn」(iSCSI修飾名)(<http://www.ietf.org/rfc/rfc3720.txt>)を参照してください。
  - d. より説明的な名前にしたい場合は、さまざまなターゲットで一意であれば、識別子を変更できます。
  - e. [追加]をクリックします。
  - f. [パス]にデバイス名を入力し、[Scsiid]を使用します。
  - g. [次へ]を2回クリックします。
6. 警告ボックスで[はい]を選択して確認します。
7. 環境設定ファイル `/etc/iscsi/iscsid.conf`を開き、パラメータ `node.startup`を `automatic` に変更します。

次の手順に従って、iSCSIイニシエータを設定します。

#### 手順 21.5: iSCSIイニシエータを設定する

1. YaSTを実行し、[ネットワークサービス] > [iSCSIイニシエータ]の順にクリックします。
2. コンピュータがブートするたびに、iSCSIイニシエータを起動したい場合は、[ブート時]を選択し、そうでない場合は、[手動]を選択します。

3. [検出]タブに切り替え、[検出] ボタンをクリックします。
4. 自分のIPアドレスとiSCSIターゲットのポートを追加します(手順21.4「iSCSIターゲット(SAN上)を設定する」参照)。通常は、ポートを既定のままにし、デフォルト値を使用できます。
5. 認証を使用する場合は、受信および送信用のユーザ名およびパスワードを挿入します。そうでない場合は、[認証なし]を選択します。
6. [次へ]を選択します。検出された接続が一覧されます。
7. [完了]をクリックして続行します。
8. シェルを開いて、rootとしてログインします。
9. iSCSIイニシエータが正常に起動しているかどうかテストします。

```
root # iscsiadm -m discovery -t st -p 192.168.3.100
192.168.3.100:3260,1 iqn.2010-03.de.jupiter:san1
```

10. セッションを確立します。

```
root # iscsiadm -m node -l -p 192.168.3.100 -T iqn.2010-03.de.jupiter:san1
Logging in to [iface: default, target: iqn.2010-03.de.jupiter:san1, portal:
192.168.3.100,3260]
Login to [iface: default, target: iqn.2010-03.de.jupiter:san1, portal:
192.168.3.100,3260]: successful
```

lsscsi でデバイス名を表示します。

```
...
[4:0:0:2] disk IET ... 0 /dev/sdd
[5:0:0:1] disk IET ... 0 /dev/sde
```

3番目の列に IET を含むエントリを捜します。この場合、該当するデバイスは、/dev/sdd と /dev/sde です。

#### 手順 21.6: LVM2ボリュームグループを作成する

1. 手順21.5「iSCSIイニシエータを設定する」のiSCSIイニシエータを実行したノードの1つで、root シェルを開きます。
2. ディスク /dev/sdd および /dev/sde でコマンド pvcreate を使用して、LVM2用に物理ボリュームを準備します。

```
root # pvcreate /dev/sdd
root # pvcreate /dev/sde
```

3. 両方のディスク上でクラスタ対応のボリュームグループを作成します。

```
root # vgcreate --clustered y clustervg /dev/sdd /dev/sde
```

4. 必要に応じて、論理ボリュームを作成します。

```
root # lvcreate -m1 --name clusterlv --size 500M clustervg
```

5. 物理ボリュームを `pvdisplay` で確認します。

```
--- Physical volume ---
PV Name /dev/sdd
VG Name clustervg
PV Size 509,88 MB / not usable 1,88 MB
Allocatable yes
PE Size (KByte) 4096
Total PE 127
Free PE 127
Allocated PE 0
PV UUID 52okH4-nv3z-2AUL-GhAN-8DAZ-GMtU-Xrn9Kh

--- Physical volume ---
PV Name /dev/sde
VG Name clustervg
PV Size 509,84 MB / not usable 1,84 MB
Allocatable yes
PE Size (KByte) 4096
Total PE 127
Free PE 127
Allocated PE 0
PV UUID Ouj3Xm-AI58-1xB1-mWm2-xn51-agM2-0UuHFC
```

6. ボリュームグループを `vgdisplay` で確認します。

```
--- Volume group ---
VG Name clustervg
System ID
Format lvm2
Metadata Areas 2
Metadata Sequence No 1
VG Access read/write
VG Status resizable
Clustered yes
Shared no
MAX LV 0
Cur LV 0
Open LV 0
Max PV 0
Cur PV 2
Act PV 2
```

|                 |                                        |
|-----------------|----------------------------------------|
| VG Size         | 1016,00 MB                             |
| PE Size         | 4,00 MB                                |
| Total PE        | 254                                    |
| Alloc PE / Size | 0 / 0                                  |
| Free PE / Size  | 254 / 1016,00 MB                       |
| VG UUID         | UCyWw8-2jqV-enuT-KH4d-NXQI-JhH3-J24anD |

ボリュームを作成してリソースを起動すると、`/dev/dm-*`という名前で新しいデバイスが作成されています。LVM2リソースの上でクラスタ化されたファイルシステム(たとえば、OCFS)を使用することをお勧めします。詳細については、「[第18章「OCFS2」](#)」を参照してください。

## 21.2.4 シナリオ - DRBDを使用するcLVM

市、国、または大陸の各所にデータセンターが分散している場合は、次のシナリオを使用できます。

手順 21.7: DRBDでクラスタ対応ボリュームグループを作成する

### 1. プライマリ/プライマリDRBDリソースを作成する

- まず、[手順20.1「DRBDの手動設定」](#)の説明に従って、DRBDデバイスをプライマリ/セカンダリとしてセットアップします。ディスクの状態が両方のノードで `up-to-date` であることを確認します。`drbdadm status` を使用してこれをチェックします。
- 次のオプションを環境設定ファイル(通常は、`/etc/drbd.d/r0.res`)に追加します。

```
resource r0 {
 net {
 allow-two-primaries;
 }
 ...
}
```

- 変更した設定ファイルをもう一方のノードにコピーします。たとえば、次のように指定します。

```
root # scp /etc/drbd.d/r0.res venus:/etc/drbd.d/
```

- 両方のノードで、次のコマンドを実行します。

```
root # drbdadm disconnect r0
root # drbdadm connect r0
root # drbdadm primary r0
```

- ノードのステータスをチェックします。

```
root # drbdadm status r0
```

2. clvmdリソースをペースメーカーの環境設定でクローンとして保存し、DLMクローンリソースに依存させます。詳細については、[手順21.1「DLMリソースを作成する」](#)を参照してください。次に進む前に、クラスタでこれらのリソースが正しく機動していることを確認してください。`crm status`またはWebインタフェースを使用して、実行中のサービスを確認できます。
3. `pvccreate` コマンドで、LVM2用に物理ボリュームを準備します。たとえば、`/dev/drbd_r0` デバイスでは、コマンドは次のようになります。

```
root # pvccreate /dev/drbd_r0
```

4. クラスタ対応のボリュームグループを作成します。

```
root # vgcreate --clustered y myclusterfs /dev/drbd_r0
```

5. 必要に応じて、論理ボリュームを作成します。論理ボリュームのサイズは変更できます。たとえば、次のコマンドで、4GBの論理ボリュームを作成します。

```
root # lvcreate -m1 --name testlv -L 4G myclusterfs
```

6. VG内の論理ボリュームは、ファイルシステムのマウントまたはraw用として使用できるようになりました。論理ボリュームを使用しているサービスにコロケーションのための正しい依存性があることを確認し、VGをアクティブ化したら論理ボリュームの順序付けを行います。

このような設定手順を終了すると、LVM2の環境設定は他のスタンドアロンワークステーションと同様に行えます。

## 21.3 有効なLVM2デバイスの明示的な設定

複数のデバイスが同じ物理ボリュームの署名を共有していると思われる場合(マルチパスデバイスやdrbdなどのように)、LVM2がPVを走査するデバイスを明示的に設定しておくことをお勧めします。

たとえばコマンド `vgcreate` がミラーブロックデバイスの代わりに物理デバイスを使用すると、DRBDは混乱してしまい、DRBDのスプリットブレイン状態が発生する場合があります。

LVM2用の単一のデバイスを非アクティブ化するには、次の手順に従います。

1. ファイル `/etc/lvm/lvm.conf` を編集し、`filter` から始まる行を検索します。
2. そこに記載されているパターンは正規表現として処理されます。冒頭の「a」は走査にデバイスパターンを受け入れることを、冒頭の「r」はそのデバイスパターンのデバイスを拒否することを意味します。
3. `/dev/sdb1` という名前のデバイスを削除するには、次の表現をフィルタルールに追加します。

```
"r|^/dev/sdb1$|"
```

完全なフィルタ行は次のようになります。

```
filter = ["r|^/dev/sdb1$|", "r|/dev/.*/by-path/.*/", "r|/dev/.*/by-id/.*/",
 "a/.*/"]
```

DRBDとMPIOデバイスは受け入れ、その他のすべてのデバイスは拒否するフィルタ行は次のようになります。

```
filter = ["a|/dev/drbd.*|", "a|/dev/.*/by-id/dm-uuid-mpath-.*/", "r/.*/"]
```

4. 環境設定ファイルを書き込み、すべてのクラスタノードにコピーします。

## 21.4 詳細

詳細な情報は、<http://www.clusterlabs.org/wiki/Help:Contents> にあるPacemakerメーリングリストから取得できます。

cLVMのFAQのオフィシャルサイトは<http://sources.redhat.com/cluster/wiki/FAQ/CLVM> です。



## 22 クラスタマルチデバイス(Cluster MD)

クラスタマルチデバイス(Cluster MD)は、クラスタ用のソフトウェアベースのRAIDストレージソリューションです。Cluster MDでは、クラスタにRAID1ミラーリングの冗長性を提供します。現在、RAID1のみがサポートされています。この章では、Cluster MDの作成および使用方法を示します。

### 22.1 概念の概要

Cluster MDは、クラスタ環境内のRAID1の使用をサポートします。Cluster MDで使用するディスクまたはデバイスは、各ノードからアクセスされます。Cluster MDの一方のデバイスに障害が発生した場合、実行時に他方のデバイスに置き換えることができ、同じ量の冗長性を提供するために再同期されます。Cluster MDでは、調整とメッセージングのためにCorosyncと分散ロックマネージャ(DLM)が必要です。

Cluster MDデバイスは、他の通常のMDデバイスのようにブート時に自動的に開始されません。クラスタ化されたデバイスはリソースエージェントを使用して開始し、DLMリソースが開始されていることを確認する必要があります。

### 22.2 クラスタ化されたMD RAIDデバイスの作成

#### 要件

- Pacemakerを使用して実行中のクラスタ。
- DLMのリソースエージェント(DLMの設定方法については、[手順17.1「DLMのベースグループの設定」](#)を参照してください)。
- 少なくとも2つの共有ディスクデバイス。デバイス障害の場合は自動的にフェールオーバーするスペアとして追加のデバイスを使用できます。
- インストール済みパッケージ `cluster-md-kmp-default`。

1. クラスタの各ノードでDLMリソースが稼動していることを確認し、リソース状態を確認するには、次のコマンドを実行します。

```
root # crm_resource -r dlm -W
```

## 2. Cluster MDデバイスを作成します。

- 既存の通常のRAIDデバイスがない場合、次のコマンドを実行し、DLMリソースが稼動しているノードでCluster MDデバイスを作成します。

```
root # mdadm --create /dev/md0 --bitmap=clustered \
--metadata=1.2 --raid-devices=2 --level=mirror /dev/sda /dev/sdb
```

Cluster MDはバージョン1.2のメタデータでのみ動作します。このため、`--metadata` オプションを使用してバージョンを指定することが推奨されます。他の役立つオプションについては、`mdadm` のマニュアルページを参照してください。`/proc/mdstat` で再同期の進捗状況を監視します。

- 既存の通常のRAIDがすでにある場合は、最初に既存のビットマップをクリアしてからクラスタ化されたビットマップを作成します。

```
root # mdadm --grow /dev/mdX --bitmap=none
root # mdadm --grow /dev/mdX --bitmap=clustered
```

- オプションで、自動フェールオーバーのためのスペアデバイスを使用してCluster MDデバイスを作成するには、1つのクラスタノード上で次のコマンドを実行します。

```
root # mdadm --create /dev/md0 --bitmap=clustered --raid-devices=2 \
--level=mirror --spare-devices=1 /dev/sda /dev/sdb /dev/sdc --
metadata=1.2
```

## 3. UUIDおよび関連するmdパスを取得します。

```
root # mdadm --detail --scan
```

このUUIDはスーパーブロックに保存されているUUIDと一致する必要があります。UUIDの詳細については、`mdadm.conf` のマニュアルページを参照してください。

4. `/etc/mdadm.conf` を開いて、mdデバイス名とそのデバイス名に関連付けられているデバイスを追加します。前のステップのUUIDを使用します。

```
DEVICE /dev/sda /dev/sdb
ARRAY /dev/md0 UUID=1d70f103:49740ef1:af2afce5:fcf6a489
```

5. Csync2の設定ファイル `/etc/csync2/csync2.cfg` を開き、`/etc/mdadm.conf` を追加します。

```
group ha_group
{
 # ... list of files pruned ...
 include /etc/mdadm.conf
}
```

## 22.3 リソースエージェントの設定

CRMリソースを次のように設定します。

1. Raid1 プリミティブを作成します。

```
crm(live)configure# primitive raider Raid1 \
 params raidconf="/etc/mdadm.conf" raiddev=/dev/md0 \
 force_clones=true \
 op monitor timeout=20s interval=10 \
 op start timeout=20s interval=0 \
 op stop timeout=20s interval=0
```

2. raider リソースをDLM用に作成したストレージのベースグループに追加します。

```
crm(live)configure# modgroup g-storage add raider
```

add サブコマンドは、デフォルトで新しいグループメンバーを追加します。

まだ実行されていない場合は、g-storage グループのクローンを作成して、すべてのノードで実行できるようにします。

```
crm(live)configure# clone cl-storage g-storage \
 meta interleave=true target-role=Started
```

3. show で変更内容をレビューします。
4. すべて正しいと思われる場合は、commit で変更を送信します。

## 22.4 デバイスの追加

デバイスを既存のアクティブなCluster MDデバイスに追加するには、最初にそのデバイスが各ノード上で「表示可能」であることを、コマンドを実行して確認します(cat /proc/mdstat)。デバイスが表示できない場合、コマンドは失敗します。

1つのクラスタノード上で次のコマンドを使用します。

```
root # mdadm --manage /dev/md0 --add /dev/sdc
```

追加された新しいデバイスの動作は、Cluster MDデバイスの状態によって異なります。

- ミラーリングされたデバイスの1つのみがアクティブである場合、新しいデバイスはミラーリングされたデバイスの2番目のデバイスになり、回復処理が開始されます。
- Cluster MDデバイスの両方のデバイスがアクティブな場合、新たに追加されたデバイスはスペアデバイスになります。

## 22.5 一時的に障害が発生したデバイスの再追加

多くの場合、障害は一時的なもので、単一ノードに限定されます。任意のノードでI/O処理中に障害が発生した場合、クラスタ全体のデバイスがfailedとマーク付けされます。

たとえば、あるノードでケーブル障害が発生した場合などに、このようになることがあります。問題を修正した後で、デバイスを再追加できます。新しいパーツを追加してデバイス全体を同期するのではなく、古いパーツのみが同期されます。

デバイスを再追加するには、1つのクラスタノード上で次のコマンドを実行します。

```
root # mdadm --manage /dev/md0 --re-add /dev/sdb
```

## 22.6 デバイスの削除

置き換えるために実行時にデバイスを削除する前に、次の操作を実行します。

1. `/proc/mdstat` をイントロスペクトしてデバイスに障害が発生しているか確認します。デバイスの前にある (F) を探します。
2. 1つのクラスタノード上で次のコマンドを実行して、デバイスに障害発生させます。

```
root # mdadm --manage /dev/md0 --fail /dev/sda
```

3. 1つのクラスタノード上で次のコマンドを実行して障害が発生したデバイスを削除します。

```
root # mdadm --manage /dev/md0 --remove /dev/sda
```

## 23 Sambaクラスタリング

クラスタ対応のSambaサーバは、異種混合ネットワークにHigh Availabilityソリューションを提供します。この章では、背景情報とクラスタ対応Sambaサーバの設定方法を説明します。

### 23.1 概念の概要

TDB (Trivial Database)は、長年にわたって、Sambaによって使用されてきました。TDBでは、複数のアプリケーションが同時に書き込むことができます。すべての書き込み操作を正常に実行し、互いに衝突させないため、TDBは、内部ロッキングメカニズムを使用しています。

CTDB (Cluster Trivial Database)は、既存のTDBの小規模な拡張です。CTDBは、プロジェクトによって、「一時データの保存のために、Sambaなどのプロジェクトによって使用されるTDBデータベースのクラスタ実装」として説明されています。

各クラスタノードは、ローカルCTDBデーモンを実行します。Sambaは、そのTDBに直接書き込むのではなく、そのローカルCTDBデーモンと通信します。それらのデーモンは、ネットワークを介してメタデータを交換しますが、実際の読み取り/書き込み操作は、高速ストレージでローカルコピー上で行われます。CTDBの概念は、[図23.1「CTDBクラスタの構造」](#)に表示されています。



#### 注記: Samba専用CTDB

CTDBリソースエージェントの現在の実装では、Sambaの管理のためだけにCTDBを設定します。他の機能(IPフェールオーバーなど)はすべて、Pacemakerで設定する必要があります。

CTDBは、完全に同種のクラスタに関してのみサポートされます。たとえば、クラスタのすべてのノードが同じアーキテクチャを持つ必要があります。x86とAMD64/Intel 64を混合することはできません。

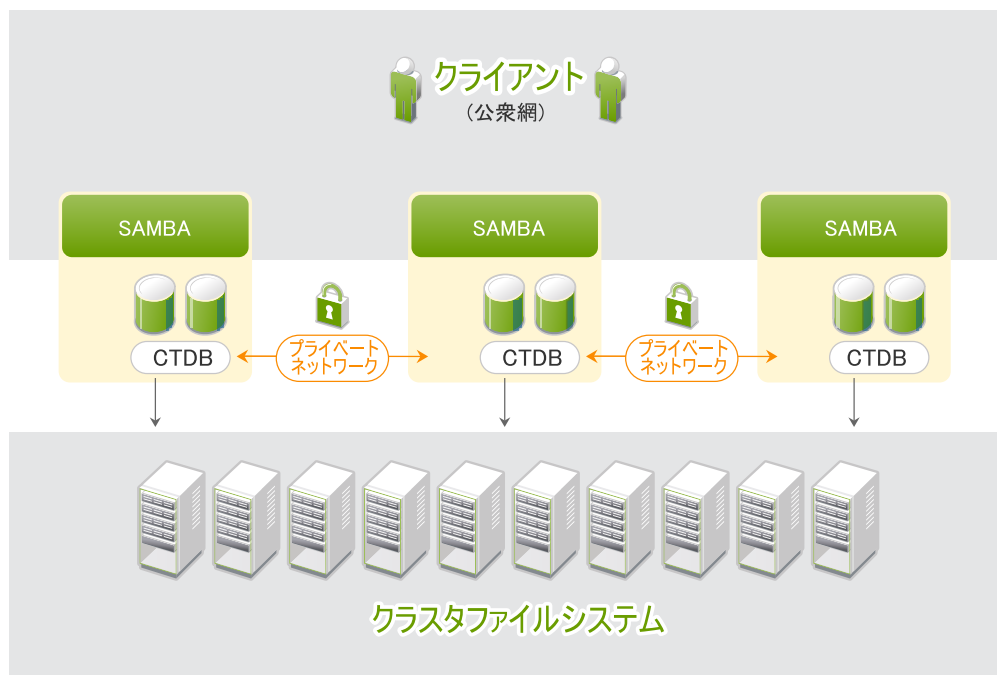


図 23.1: CTDBクラスタの構造

クラスタ対応Sambaサーバは、一定のデータを共有する必要があります。

- UnixのユーザとグループIDをWindowsのユーザとグループに関連付けるマッピングテーブル。
- ユーザデータベースをすべてのノード間で同期する必要があります。
- Windowsドメイン内のメンバサーバの参加情報をすべてのノードで利用できる必要があります。
- メタデータをすべてのノードで利用できる必要があります(アクティブSMBセッション、共有接続、各ロックなど)。

クラスタ対応SambaサーバがN+1ノードを持っている場合、Nノードだけのサーバより高速になることを目的としています。1つのノードは、クラスタ非対応のSambaサーバより遅くなることはありません。

## 23.2 基本的な設定



### 注記: 変更された設定ファイル

CTDBリソースエージェントは、自動的に `/etc/sysconfig/ctdb` を変更します。`crm ra info CTDB` を使用して、CTDBリソースに指定できるすべてのパラメータを一覧表示してください。

クラスタ対応Sambaサーバをセットアップするには、次の手順に従います。

1. クラスタを準備します。

- a. 次のパッケージがインストールされていることを確認してから進んでください: `ctdb`、`tdb-tools`、および `samba` (`smb` および `nmb` リソースに必要)。
- b. このガイドの [パートII「設定および管理」](#) で説明されているように、クラスタ (Pacemaker、OCFS2) を設定します。
- c. OCFS2などの共有ファイルシステムを設定し、マウントします(たとえば、`/srv/clusterfs` にマウント)。詳細については、[第18章「OCFS2」](#) を参照してください。
- d. POSIX ACLをオンにする場合は、それを有効にします。

- 新しいOCFS2ファイルシステムの場合は、次のコマンドを使用します。

```
root # mkfs.ocfs2 --fs-features=xattr ...
```

- 既存のOCFS2ファイルシステムの場合は、次のコマンドを使用します。

```
root # tuneefs.ocfs2 --fs-feature=xattr DEVICE
```

ファイルシステムリソースには、必ず、`acl` オプションを指定します。次のように、`crm` シェルを使用します。

```
crm(live)configure# primitive ocfs2-3 ocf:heartbeat:Filesystem params
options="acl" ...
```

- e. `ctdb`、`smb`、`nmb` の各サービスが無効になるようにします。

```
root # systemctl disable ctdb
root # systemctl disable smb
root # systemctl disable nmb
```

- f. すべてのノードのファイアウォールのポート `4379` を開きます。これは、CTDBが他のクラスタノードと通信するために必要です。

2. 共有ファイルシステムにCTDBロックのディレクトリを作成します。

```
root # mkdir -p /srv/clusterfs/samba/
```

3. `/etc/ctdb/nodes` に、クラスタ内の各ノードの全プライベートIPアドレスを含むすべてのノードを挿入します。

```
192.168.1.10
```

4. Sambaを設定します。`/etc/samba/smb.conf`の`[global]`セクションに次の行を追加します。「CTDB-SERVER」の代わりに、選択したホスト名を使用します(クラスタ内のすべてのノードは、この名前を持つ1つの大きなノードとして表示されます)。

```
[global]
...
settings applicable for all CTDB deployments
netbios name = CTDB-SERVER
clustering = yes
idmap config * : backend = tdb2
passdb backend = tdbsam
ctdbd socket = /var/lib/ctdb/ctdb.socket
settings necessary for CTDB on OCFS2
fileid:algorithm = fsid
vfs objects = fileid
...
```

5. `csync2`を使用して、設定ファイルをすべてのノードにコピーします。

```
root # csync2 -xv
```

詳細については、[手順4.6「Csync2による設定ファイルの同期」](#)を参照してください。

6. CTDBリソースをクラスタに追加します。

```
root # crm configure
crm(live)configure# primitive ctdb ocf:heartbeat:CTDB params \
 ctdb_manages_winbind="false" \
 ctdb_manages_samba="false" \
 ctdb_recovery_lock="/srv/clusterfs/samba/ctdb.lock" \
 ctdb_socket="/var/lib/ctdb/ctdb.socket" \
 op monitor interval="10" timeout="20" \
 op start interval="0" timeout="90" \
 op stop interval="0" timeout="100"
crm(live)configure# primitive nmb systemd:nmb \
 op start timeout="60" interval="0" \
 op stop timeout="60" interval="0" \
 op monitor interval="60" timeout="60"
crm(live)configure# primitive smb systemd:smb \
 op start timeout="60" interval="0" \
 op stop timeout="60" interval="0" \
 op monitor interval="60" timeout="60"
crm(live)configure# group g-ctdb ctdb nmb smb
crm(live)configure# clone cl-ctdb g-ctdb meta interleave="true"
crm(live)configure# colocation col-ctdb-with-clusterfs inf: cl-ctdb cl-clusterfs
crm(live)configure# order o-clusterfs-then-ctdb inf: cl-clusterfs cl-ctdb
crm(live)configure# commit
```



7. クラスタ対応のIPアドレスを追加します。

```
crm(live)configure# primitive ip ocf:heartbeat:IPaddr2 params ip=192.168.2.222 \
 unique_clone_address="true" \
 op monitor interval="60" \
 meta resource-stickiness="0"
crm(live)configure# clone cl-ip ip \
 meta interleave="true" clone-node-max="2" globally-unique="true"
crm(live)configure# colocation col-ip-with-ctdb 0: cl-ip cl-ctdb
crm(live)configure# order o-ip-then-ctdb 0: cl-ip cl-ctdb
crm(live)configure# commit
```

`unique_clone_address` が `true` に設定されている場合、IPaddr2リソースエージェントはクローンIDを指定のアドレスに追加し、3つの異なるIPアドレスを設定します。これらは通常必要とされませんが、負荷分散に役立ちます。この項目の詳細については、[14.2項「Linux仮想サーバによる負荷分散の設定」](#)を参照してください。

8. 変更をコミットします。

```
crm(live)configure# commit
```

9. 結果を確認します。

```
root # crm status
Clone Set: cl-storage [dlm]
 Started: [factory-1]
 Stopped: [factory-0]
Clone Set: cl-clusterfs [clusterfs]
 Started: [factory-1]
 Stopped: [factory-0]
Clone Set: cl-ctdb [g-ctdb]
 Started: [factory-1]
 Started: [factory-0]
Clone Set: cl-ip [ip] (unique)
 ip:0 (ocf:heartbeat:IPaddr2): Started factory-0
 ip:1 (ocf:heartbeat:IPaddr2): Started factory-1
```

10. クライアントコンピュータからテストを行います。次のコマンドをLinuxクライアントで実行して、システムからファイルをコピーしたり、システムにファイルをコピーできるかどうか確認します。

```
root # smbclient //192.168.2.222/myshare
```

## 23.3 Active Directoryドメインへの追加

Active Directory (AD)は、Windowsサーバシステムのディレクトリサービスです。

次の手順は、CTDBクラスタをActive Directoryドメインに追加する方法を概説しています。

1. 手順23.1「クラスタ対応Sambaサーバの基本セットアップ」の説明に従って、CTDBリソースを作成します。
2. `samba-winbind` パッケージをインストールします。
3. `winbind` サービスを無効にします。

```
root # systemctl disable winbind
```

4. `winbind` クラスタリソースを定義します。

```
root # crm configure
crm(live)configure# primitive winbind systemd:winbind \
 op start timeout="60" interval="0" \
 op stop timeout="60" interval="0" \
 op monitor interval="60" timeout="60"
crm(live)configure# commit
```

5. `g-ctdb` グループを編集して、`nmb` と `smb` リソースの間に `winbind` を挿入します。

```
crm(live)configure# edit g-ctdb
```

`:w` (`vim`)でエディタを保存して閉じます。

6. Active Directoryドメインのセットアップ方法については、Windows Serverのマニュアルを参照してください。この例では、次のパラメータを使用します。

|                     |                            |
|---------------------|----------------------------|
| ADおよびDNSサーバ         | win2k3.2k3test.example.com |
| ADドメイン              | 2k3test.example.com        |
| クラスタADメンバーのNETBIOS名 | CTDB-SERVER                |

7. 手順23.2「Active Directoryへの参加」

最後に、クラスタをActive Directoryサーバに参加させます。

手順 23.2: ACTIVE DIRECTORYへの参加

1. 次のファイルが、すべてのクラスタホストにインストールされるように、Csync2設定に含まれていることを確認します。

```
/etc/samba/smb.conf
```

```
/etc/security/pam_winbind.conf
/etc/krb5.conf
/etc/nsswitch.conf
/etc/security/pam_mount.conf.xml
/etc/pam.d/common-session
```

この作業には、YaSTの[Csync2の設定]モジュールを使用することもできます。[4.5項「すべてのノードへの設定の転送」](#)を参照してください。

2. YaSTを実行し、[ネットワークサービス]エントリから[Windowsドメインメンバーシップ]モジュールを開きます。
3. ドメインまたはワークグループの設定を入力して、[OK]をクリックして終了します。

## 23.4 クラスタ対応Sambaのデバッグとテスト

クライアント対応Sambaサーバのデバッグには、次のツールを使用できます。これらのツールは、さまざまなレベルで動作します。

### `ctdb_diagnostics`

このツールを実行して、クラスタ対応Sambaサーバを診断します。詳細なデバッグメッセージが出力されるので、発生している問題を追跡するのに役立ちます。

`ctdb_diagnostics` コマンドは、次のファイルを検索します。これらのファイルは、すべてのノードで利用できる必要があります。

```
/etc/krb5.conf
/etc/hosts
/etc/ctdb/nodes
/etc/sysconfig/ctdb
/etc/resolv.conf
/etc/nsswitch.conf
/etc/sysctl.conf
/etc/samba/smb.conf
/etc/fstab
/etc/multipath.conf
/etc/pam.d/system-auth
/etc/sysconfig/nfs
/etc/exports
/etc/vsftpd/vsftpd.conf
```

`/etc/ctdb/public_addresses` ファイルと `/etc/ctdb/static-routes` ファイルが存在する場合は、それらもチェックされます。

## ping\_pong

`ping_pong` では、ファイルシステムがCTDBに適合しているかどうかチェックできます。このコマンドは、クラスタファイルシステムの一定のテスト(コヒーレンスやパフォーマンスなどのテスト)を実行して([http://wiki.samba.org/index.php/Ping\\_pong](http://wiki.samba.org/index.php/Ping_pong) 参照)、高負荷の状況下におけるクラスタの動作を示す情報を提供します。

## send\_arp ツールおよび SendArp リソースエージェント

`SendArp` リソースエージェントは、`/usr/lib/heartbeat/send_arp` (または `/usr/lib64/heartbeat/send_arp`) にあります。`send_arp` ツールは Gratuitous ARP (余計なアドレス解決プロトコル) パケットを送信し、他のマシンのARPテーブルを更新するために使用できます。これは、フェールオーバープロセス後の通信問題の識別に役立ちます。Sambaのクラスタ化されたIPアドレスを表示しているのに関わらず、ノードに接続またはpingできない場合は、`send_arp` コマンドを使用して、ノードはARPテーブルの更新のみが必要であるのかをテストします。詳細については、[http://wiki.wireshark.org/Gratuitous\\_ARP](http://wiki.wireshark.org/Gratuitous_ARP) を参照してください。

クラスタファイルシステムの特定の側面をテストするには、次の手順に従います。

### 手順 23.3: クラスタファイルシステムのコヒーレンスとパフォーマンスをテストする

1. 1つのノードで `ping_pong` コマンドを開始します。プレースホルダ `N` はノード数+1で置き換えます。共有ストレージでは `ABSPATH/data.txt` ファイルが使用可能で、すべてのノード上でアクセスできます (`ABSPATH` は絶対パスを示しています)。

```
ping_pong ABSPATH/data.txt N
```

1つのノードでだけ実行しているので、ロックングレートは非常に高いと予想してください。プログラムがロックングレートを出力しない場合は、クラスタファイルシステムを置き換えます。

2. 同じパラメータを使用して、別のノードで `ping_pong` の2つ目のコピーを開始します。ロックングレートが大幅に下がることを予想できます。使用しているクラスタファイルシステムに次のどれかが当てはまる場合は、クラスタファイルシステムを置き換えます。
  - `ping_pong` がロックングレート(秒単位)を出力しない。
  - 2つのインスタンスのロックングレートがほぼ同じではない。
  - 2つ目のインスタンスの開始後にロックングレートが下がらなかった。
3. `ping_pong` の3つ目のコピーを開始します。もう1つノードを追加し、ロックングレートの変化に注目します。

4. `ping_pong` コマンドを1つずつ終了させます。単一ノードの状態に戻るまで、ロッキングレートの増加が観察されるはずです。予想したような振る舞いが見られなかった場合には、[第18章「OCFS2」](#)に記されている詳細を参照してください。

## 23.5 その他の情報

- [http://wiki.samba.org/index.php/CTDB\\_Setup](http://wiki.samba.org/index.php/CTDB_Setup) 
- <http://ctdb.samba.org> 
- [http://wiki.samba.org/index.php/Samba\\_%26\\_Clustering](http://wiki.samba.org/index.php/Samba_%26_Clustering) 

## 24 Relax-and-Recover (Rear)による障害復旧

Relax-and-Recover (旧称「ReaR」。この章ではRearと略記)は、システム管理者による使用を意図した障害復旧フレームワークです。Rearは、障害発生時に保護対象となる特定の運用環境に合わせて調整する必要があるBashスクリプトのコレクションです。

特別な設定を必要とせずにそのまま使用できる障害復旧ソリューションはありません。したがって、障害が発生する前に準備しておくことが不可欠です。

### 24.1 概念の概要

以降のセクションでは、障害復旧の一般的な概念を述べ、Rearを使用した復旧を実現するために必要となる基本的な手順について説明します。また、Rearの要件に関するいくつかの指針、知っておくべき制限事項、およびシナリオとバックアップツールについても紹介します。



#### 注記: Rearについて

Rearの複雑な機能を理解することは、ツールでの作業を意図したように行うために重要です。したがって障害が発生する前に、この章を注意深く読んで、Rearについての理解を深めてください。また、Rearの既知の制限事項を認識し、システムをあらかじめテストする必要もあります。

#### 24.1.1 障害復旧プランの作成

最悪のシナリオが発生する前に、ITインフラストラクチャに重大なリスクがあるかどうかの分析、予算の見積もり、障害復旧プランの作成などの対応策を講じます。障害復旧プランを手元に用意していない場合は、以下の手順ごとに情報を入手します。

- **リスクの分析:** インフラの確かなリスク分析を実施します。可能性のあるすべての脅威を一覧表示し、深刻度を評価します。これらの脅威が発生する可能性を判断し、優先順位を設定します。可能性と影響の簡単な分類を使用することを推奨します。
- **予算のプランニング:** 分析結果には、どのリスクが耐えうるもので、どれがビジネスにとって致命的であるかを全般的に示します。リスクを最小限にする方法およびそれに要するコストを自問して検討します。会社の規模に応じて、IT予算全体の2～15%を障害復旧に使用します。

- **障害復旧プランの作成:** チェックリストの作成、手順のテスト、優先順位の設定と割り当て、ITインフラのインベントリ調査を行います。インフラのサービスが失敗した際、問題に対処する方法を定義します。
- **テスト:** 念入りなプランを定義したら、それをテストします。最低でも1年に1度テストします。ご使用のメインITインフラと同じテストハードウェアを使用します。

### 24.1.2 障害復旧とは

運用環境に存在するシステムが、ハードウェアの損傷、誤設定、ソフトウェア上の問題など、原因がどのようなものであっても破損した場合は、システムを再作成する必要があります。再作成は、同じハードウェア上または互換性のある代替ハードウェア上で実行できます。バックアップからファイルを復元するだけでは、システムを再作成することにはなりません。システムの再作成では、パーティション、ファイルシステム、マウントポイントの面からのシステムのストレージ作成や、ブートローダの再インストールなどの作業も必要になります。

### 24.1.3 Rearによる障害復旧

システムが正常に稼働しているときに、ファイルのバックアップを作成し、復旧メディア上に復旧システムを作成します。復旧システムには、復旧インストーラが収められています。

システムが破損した場合は、損傷したハードウェアを必要に応じて交換し、復旧メディアから復旧システムをブートして復旧インストーラを起動します。復旧インストーラによるシステムの再作成では、まず、ストレージにパーティション、ファイルシステム、マウントポイントを作成し、続いてバックアップからファイルを復元します。最後に、ブートローダを再インストールします。

### 24.1.4 Rearの要件

Rearを使用するには、運用環境を実行するマシンおよびそれと同一のテストマシンが必要です。つまり、同一のシステムが2台以上必要になります。ここでいう「同一」とは、たとえば、ネットワークカードを、同じカーネルドライバを使用する他のネットワークカードに置き換えることができるということです。



#### **警告: 同一のドライバが必要**

運用環境のドライバと同じドライバを使用していないハードウェアコンポーネントは、Rearでは同一のコンポーネントとは見なされません。



## 24.1.5 Rearバージョンの更新

より古いバージョンのサービスパックと互換性を持つために、SUSE Linux Enterprise High Availability Extension 12 SP5には、異なるRearバージョン: 1.16 (RPMパッケージ `rear116` の一部)、1.17.2.a (`rear1172a`)、1.18a (`rear118a`)、および2.4 (`rear23a`)が付属しています。最新バージョンには、アップストリームGitHubプロジェクトからの、若干の最新の拡張が含まれます。



### 注記: 変更ログでの重要な情報の検索

バグ修正、非互換性、および他の問題に関する情報はすべて、パッケージの変更ログで検索できます。障害復旧手順を再検証する必要がある場合は、Rearのより最新のパッケージバージョンも確認することをお勧めします。

Rearには次の問題がありますので注意してください。

- UEFIシステムで障害復旧を許可するには、バージョン1.18.aおよびパッケージ `ebiso` が必要です。このバージョンのみが新しいヘルパーツール `/usr/bin/ebiso` をサポートします。このヘルパーツールは、UEFIブート可能RearシステムISOイメージの作成に使用されます。
- 特定のRearバージョンによる障害復旧手順をテスト済みで、その手順が十分に機能しているのであれば、Rearを更新しないでください。Rearパッケージをそのまま保持し、障害復旧手法を変更しないようにします。
- Rearの各バージョン更新は、インストール済みのバージョンが誤って別のバージョンに置き換えられないことがないよう、相互に意図的に衝突する別個のパッケージとして提供されています。

次の場合に、既存の障害復旧手順を完全に再検証する必要があります。

- Rearバージョンの更新ごとに。
- Rearを手動で更新する場合。
- Rearで使用するソフトウェアごとに。
- `parted`、`btrfs`、などの低レベルのシステムコンポーネントを更新する場合。

## 24.1.6 Btrfsに伴う制限事項

Btrfsを使用する場合は次の制限事項が発生します。

システムにサブボリュームは存在しても、スナップショットのサブボリュームが存在しない場合

Rearバージョン1.17.2.a以上が必要です。このバージョンは、スナップショットのサブボリュームが存在しない「通常の」Btrfsサブボリューム構造の再作成をサポートしています。





## 警告

ファイルベースのバックアップソフトウェアでは、Btrfsスナップショットのサブボリュームを通常どおりにはバックアップできず、復元することもできません。

Btrfsにはコピーオンライト機能があることから、Btrfsファイルシステム上にある最近のスナップショットサブボリュームはディスク容量をほとんど必要としません。一方、ファイルベースのバックアップソフトウェアを使用すると、これらのファイルは完全なファイルとしてバックアップされます。最終的に、これらのファイルはその本来のファイルサイズで2回バックアップされることになります。したがって、元のシステム上に以前に存在していたときと同じ状態にスナップショットを復元することができません。

### SLE12システムに適合するRear設定が必要な場合

SLE12 GA、SLE12 SP1、およびSLE12 SP2の設定には、いくつかの不適合Btrfsのデフォルト構造があります。そのため、適合するRear設定ファイルを使用することはたいへん重要です。[/usr/share/rear/conf/examples/SLE12\\*-btrfs-example.conf](#)のサンプルファイルを参照してください。

## 24.1.7 シナリオとバックアップのツール

Rearでは、ハードディスク、フラッシュディスク、DVD/CD-RなどのローカルメディアからのブートやPXEを介したブートが可能な障害復旧システム(システム固有の復旧インストーラなど)を作成できます。バックアップデータは、[例 24.1](#)に説明があるNFSなどのネットワークファイルシステムに保存できます。

Rearは、ファイルのバックアップに取って代わるツールではなく、それを補完するツールです。Rearは、汎用的な `tar` コマンドのほか、いくつかのサードパーティのバックアップツールをデフォルトでサポートしています。このようなバックアップツールとして、Tivoli Storage Manager、QNetix Galaxy、Symantec NetBackup、EMC NetWorker、HP DataProtectorなどがあります。バックアップツールとしてEMC NetWorkerを使用したRearの設定例については[例 24.2](#)を参照してください。

## 24.1.8 基本手順

障害発生時にRearを使用して効果的な復旧を実現するには、以下の基本手順を実行する必要があります。

## Rearおよびバックアップソリューションのセットアップ

この手順では、Rear設定ファイルの編集、Bashスクリプトの調整、使用するバックアップソリューションの設定などのタスクを実行します。

### 復旧インストールシステムの作成

保護対象のシステムが稼働しているときに、`rear mkbbackup` コマンドを使用して、ファイルのバックアップを作成し、システム固有のRear復旧インストーラなどの復旧システムを生成します。

### 復旧プロセスのテスト

Rearを使用して障害復旧メディアを作成した場合は、その障害復旧プロセスを必ず十分にテストしておくようにします。ここでは、運用環境を構成するハードウェアと同一のハードウェアを備えたテストマシンの使用が不可欠です。詳細については、[24.1.4項「Rearの要件」](#)を参照してください。

### 障害からの復旧

障害が発生した場合は、必要に応じて損傷したハードウェアを交換します。続いて、Rear復旧システムをブートし、`rear recover` コマンドで復旧インストーラを起動します。

## 24.2 Rearおよびバックアップソリューションのセットアップ

Rearをセットアップするには、少なくともRear設定ファイル `/etc/rear/local.conf` を編集する必要があります。さらに、Rearフレームワークを構成するBashスクリプトも必要に応じて編集します。

特に、Rearが実行する以下のタスクの定義が必要です。

- システムをUEFIを使用してブートする場合： システムをUEFIブートローダを使用してブートする場合は、パッケージ `ebiso` をインストールし、次の行を `/etc/rear/local.conf` に追加します。

```
ISO_MKISOFS_BIN=/usr/bin/ebiso
```

- ファイルをバックアップする方法および障害復旧システムを作成して保存する方法： これは、`/etc/rear/local.conf` で設定する必要があります。
- 正確な再作成を必要とする対象(パーティション、ファイルシステム、マウントポイントなど)： これは、`/etc/rear/local.conf` で定義できます(たとえば、除外対象を定義できます)。標準とは異なるシステムを再作成するには、Bashスクリプトの拡張が必要になることがあります。
- 復旧プロセスの仕組み： Rearで復旧インストーラを生成する方法の変更やRear復旧インストーラによる実行タスクとの適合を可能にするには、Bashスクリプトの編集が必要です。

Rearを設定するには、`/etc/rear/local.conf` 設定ファイルに目的のオプションを追加します(これまで使用されてきた設定ファイル `/etc/rear/sites.conf` は、パッケージから削除されました。なお、このファイルを前回のセットアップから引き継いでいる場合、Rearでは引き続きこのファイルが使用されます)。

すべてのRear設定変数とそのデフォルト値は、`/usr/share/rear/conf/default.conf` に設定されています。`/etc/rear/local.conf` などで設定されているユーザ設定に合わせたサンプルファイルのいくつか( `*example.conf` )は、`examples` サブディレクトリ下にあります。詳細については、Rearのマニュアルで該当のページを参照してください。

適合するサンプル設定ファイルをまずはテンプレートとして使用し、必要に応じて修正することで、個別の設定ファイルを作成します。いくつかのサンプル設定ファイルからさまざまなオプションをコピーし、システムに適合した特定の `/etc/rear/local.conf` ファイルにそれらをペーストします。特定の構成向けに利用できる変数の概要などが含まれるため、元のサンプル設定ファイルをそのまま使用しないでください。

Rear設定ファイルを変更した場合は、次のコマンドを実行して、その出力を確認します。

```
rear dump
```

#### 例 24.1: NFSサーバを使用したファイルバックアップの保存

Rearはさまざまなシナリオで使用できます。以下の例では、ファイルのバックアップを収めるストレージとしてNFSサーバを使用しています。

1. 『SUSE Linux Enterprise Server 12 SP5管理ガイド』(<https://documentation.suse.com/sles-12/html/SLES-all/cha-nfs.html>)の説明に従って、YaSTでNFSサーバを設定します。
2. 目的のNFSサーバの設定を `/etc/exports` ファイルで定義します。NFSサーバ上でバックアップデータの保存先とするディレクトリに、適切なマウントオプションが設定されていることを確認します。次に例を示します。

```
/srv/nfs *([...],rw,no_root_squash,...)
```

`/srv/nfs` をNFSサーバ上のバックアップデータへのパスに置き換えて、マウントオプションを調整します。バックアップデータにアクセスするには、`no_root_squash` の指定が必要になることが普通です。これは、`rear mkbackup` コマンドが `root` として実行されるためです。

3. さまざまな `BACKUP` パラメータ(設定ファイル `/etc/rear/local.conf` に記述されています)を調整して、該当のNFSサーバ上にRearからファイルのバックアップを保存できるようにします。この例は、インストールしたシステムの `/usr/share/rear/conf/examples/SLE12-*example.conf` にあります。

`tar` の代わりにサードパーティのバックアップツールを使用するには、Rear設定ファイルを適切に設定する必要があります。

以下は、EMC NetWorkerを使用する場合の設定例です。この設定スニペットを `/etc/rear/local.conf` に追加し、それぞれのセットアップに応じて調整します。

```
BACKUP=NSR
OUTPUT=ISO
BACKUP_URL=nfs://host.example.com/path/to/rear/backup
OUTPUT_URL=nfs://host.example.com/path/to/rear/backup
NSRSERVER=backupserver.example.com
RETENTION_TIME="Month"
```

## 24.3 復旧インストールシステムの作成

24.2項の説明に従ってRearを設定した後、Rear復旧インストーラを持つ復旧インストールシステムを作成したうえで、以下のコマンドを使用してファイルのバックアップを作成します。

```
rear -d -D mkbackup
```

このコマンドでは以下の手順が実行されます。

1. ターゲットシステムを分析し、ディスクのレイアウト(パーティション、ファイルシステム、マウントポイント)やブートローダに関する情報を中心として必要な情報を収集する。
2. 最初の手順で収集した情報を使用して、ブート可能な復旧システムを作成する。ここで得られるRear復旧インストーラは、障害から保護する個々のシステム専用のインストーラです。このインストーラは、この固有のシステムを再作成する目的でのみ使用できます。
3. 設定済みのバックアップツールを呼び出し、システムとユーザファイルをバックアップする。

## 24.4 復旧プロセスのテスト

復旧システムを作成した後、運用マシンと同一のハードウェアを備えたテストマシンで復旧プロセスをテストします。24.1.4項「Rearの要件」も参照してください。テストマシンの設定が適切で、メインマシンの代わりとして機能できることを確認します。



### 警告: 運用マシンと同一のハードウェア上での包括的なテスト

マシン上で障害復旧プロセスを十分にテストする必要があります。復旧手順を定期的にテストし、すべてが想定どおりに機能することを確認します。

1. 24.3項で作成した復旧システムをDVDやCDに書き込み、復旧メディアを作成します。PXEを介したネットワークブートとすることもできます。
2. 復旧メディアからテストマシンをブートします。
3. メニューから[Recover (復旧)]を選択します。
4. rootとしてログインします(パスワードは必要なし)。
5. 次のコマンドを入力して復旧インストーラを起動します。

```
rear -d -D recover
```

このプロセスでRearが実行する手順の詳細については[回復プロセス](#)を参照してください。

6. 復旧プロセスが完了した後、システムが正常に再作成されたかどうか、および運用環境で元のシステムの代替として機能するかどうかを確認します。

## 24.5 障害からの復旧

障害が発生した場合には、必要に応じて損傷したハードウェアを取り替えます。次に、[手順 24.1](#)の説明に従って、修復したマシンまたは元のシステムの代替として機能することをテスト済みの同一構成のマシンを使用して手順を進めます。

`rear recover` コマンドでは以下の手順が実行されます。

### 回復プロセス

1. ディスクのレイアウト(パーティション、ファイルシステム、およびマウントポイント)を復元する。
2. バックアップからシステムとユーザファイルを復元する。
3. ブートローダを復元する。

## 24.6 その他の情報

- [http://en.opensuse.org/SDB:Disaster\\_Recovery](http://en.opensuse.org/SDB:Disaster_Recovery) 
- `rear` マニュアルページ
- </usr/share/doc/packages/rear/README>

## IV 付録

- A   トラブルシューティング 308
- B   命名規則 318
- C   クラスタ管理ツール(コマンドライン) 319
- D   rootアクセスなしでのクラスタレポートの実行 321

## A トラブルシューティング

時として理解しにくい奇妙な問題が発生することがあります。High Availabilityでの実験を開始したときには、特にそうです。それでも、High Availabilityの内部プロセスを詳しく調べるために使用できる、いくつかのユーティリティがあります。この章では、さまざまなソリューションを推奨します。

### A.1 インストールと最初のステップ

パッケージのインストールやクラスタのオンライン化では、次のように問題をトラブルシュートします。

#### HAパッケージはインストールされているか

クラスタの構成を管理に必要なパッケージは、High Availability Extensionで使える High Availability インストールパターンに付属しています。

High Availability Extensionが各クラスタノードにSUSE Linux Enterprise Server 12 SP5の拡張としてインストールされているか、[High Availability] パターンが『インストールおよびセットアップクイックスタート』で説明するように各マシンにインストールされているか、確認します。

#### 初期設定がすべてのクラスタノードについて同一か

相互に通信するため、第4章「YaSTクラスタモジュールの使用」で説明するように、同じクラスタに属するすべてのノードは同じ bindnetaddr、mcastaddr、mcastport を使用する必要があります。

/etc/corosync/corosync.conf で設定されている通信チャンネルとオプションがすべてのクラスタノードに関して同一かどうか確認します。

暗号化通信を使用する場合は、/etc/corosync/authkey ファイルがすべてのクラスタノードで使用可能かどうかを確認します。

すべての corosync.conf 設定(nodeid 以外)が同一で、すべてのノードの authkey ファイルが同一でなければなりません。

#### ファイアウォールで mcastport による通信が許可されているか

クラスタノード間の通信に使用される mcastport がファイアウォールでブロックされている場合、ノードは相互に認識できません。第4章「YaSTクラスタモジュールの使用」と『インストールおよびセットアップクイックスタート』で説明されているように、YaSTまたはブートストラップスクリプトで初期セットアップを設定しているときに、ファイアウォール設定は通常、自動的に調整されます。



mcastportがファイアウォールでブロックされないようにするには、各ノードの /etc/sysconfig/SuSEfirewall12 の設定を確認します。または、各クラスタノードのYaSTファイアウォールモジュールを起動します。[\[許可されるサービス\]](#) > [\[の詳細\]](#) をクリックして、mcastportを許可された[UDPポート]のリストに追加し、変更を確定します。

PacemakerとCorosyncが各クラスタノードで開始されているか

通常、Pacemakerを開始すると、Corosyncサービスも開始します。両方のサービスが実行されているかどうかを確認するには、次のコマンドを実行します。

```
root # systemctl status pacemaker corosync
```

両方のサービスが実行されていない場合は、次のコマンドを実行して開始します。

```
root # systemctl start pacemaker
```

## A.2 ログ記録

ログファイルはどこにあるか

Pacemakerログファイルの場合は、/etc/corosync/corosync.conf の logging セクションで指定されている設定を参照してください。ここで指定したログファイルをPacemakerで無視する場合は、Pacemaker独自の設定ファイル /etc/sysconfig/pacemaker のログ記録設定を確認してください。PCMK\_logfile がそこで設定されている場合、Pacemakerはこのパラメータで定義したパスを使用します。

すべての関連ログファイルを表示するクラスタ全体のレポートが必要な場合は、[詳細についてすべてのクラスタノードの分析を含むレポートを作成するにはどうしたらよいですか。](#)を参照してください。

監視を有効にしているのに、ログファイルに監視操作の記録が残っていないのはなぜですか。

lrmd デーモンは、エラーが発生しない限り、複数の監視操作はログに記録しません。複数の監視操作をすべてログ記録すると、多量のノイズが発生してしまいます。そのため、複数の監視操作は、1時間に1度だけ記録されます。

failed メッセージだけが出ました。詳細情報を取得できますか。

コマンドに --verbose パラメータを追加してください。これを複数回行くと、デバッグ出力が非常に詳細になります。役立つヒントについては、ログ記録データ(sudo journalctl -n)を参照してください。

ノードとリソースすべての概要を確認するにはどうしたらよいですか。

crm\_mon コマンドを使用してください。次のコマンドは、リソース操作履歴(-o オプション)と非アクティブなリソース(-r)を表示します。



```
root # crm_mon -o -r
```

表示内容は、ステータスが変わると、更新されます(これをキャンセルするには、**Ctrl-C** を押します)。次に例を示します

例 A.1: 停止されたリソース

```
Last updated: Fri Aug 15 10:42:08 2014
Last change: Fri Aug 15 10:32:19 2014
Stack: corosync
Current DC: bob (175704619) - partition with quorum
Version: 1.1.12-ad083a8
2 Nodes configured
3 Resources configured

Online: [alice bob]

Full list of resources:

my_ipaddress (ocf:heartbeat:Dummy): Started bob
my_filesystem (ocf:heartbeat:Dummy): Stopped
my_webserver (ocf:heartbeat:Dummy): Stopped

Operations:
* Node bob:
 my_ipaddress: migration-threshold=3
 + (14) start: rc=0 (ok)
 + (15) monitor: interval=10000ms rc=0 (ok)
* Node alice:
```

『Explained (Pacemaker)』PDF (<http://www.clusterlabs.org/doc/> から入手可能)では、「How are OCF Return Codes Interpreted?」セクションで3つの異なる復元タイプを説明しています。

ログはどのように表示しますか。

クラスタで発生している現象をより詳しく表示するには、次のコマンドを使用します。

```
root # crm history log [NODE]
```

NODE は、調べたいノードに置き換えるか、空のままにします。詳細については、[A.5項「履歴」](#)を参照してください。

## A.3 リソース

リソースはどのようにクリーンアップしますか。

次のコマンドを使用してください。

```
root # crm resource list
crm resource cleanup rscid [node]
```

ノードを指定しないと、すべてのノードでリソースがクリーンアップされます。詳細については、[8.5.3項「リソースのクリーンアップ」](#)を参照してください。

現在既知のリソースを一覧表示するにはどうしたらよいですか。

コマンド `crm_resource list` を使用して、現在のリソースの情報を表示できます。

リソースを設定しましたが、いつも失敗します。なぜですか。

OCFスクリプトを確認するには、たとえば、次の `ocf-tester` コマンドを使用します。

```
ocf-tester -n ip1 -o ip=YOUR_IP_ADDRESS \
/usr/lib/ocf/resource.d/heartbeat/IPaddr
```

パラメータを増やすには、`-o` を複数回使用します。必須パラメータとオプションパラメータのリストは、`crm ra info AGENT` の実行によって取得できます。たとえば、次のようにします。

```
root # crm ra info ocf:heartbeat:IPaddr
```

`ocf-tester` を実行する場合は、その前に、リソースがクラスタで管理されていないことを確認してください。

リソースがフェールオーバーせず、エラーが出ないのはなぜですか。

終端ノードは `unclean` (アンクリーン) と考えられる場合があります。その場合には、それをフェンシングする必要があります。STONITHリソースが動作していない、または存在しない場合、残りのノードはフェンシングが実行されるのを待機することになります。フェンシングのタイムアウトは通常長いので、問題の兆候がはっきりと現れるまでには(仮に現れたとしても)、かなり長い時間がかかることがあります。

さらに別の可能性としては、単にこのノードでのリソースの実行が許可されていないという場合があります。このことは、過去にエラーが発生し、それが正しく「解決」されていないために生じることがあります。または、以前に行った管理上の操作が原因である場合もあります。つまり、負のスコアを持つ場所の制約のためです。そのような場所の制約は、たとえば、`crm resource migrate` コマンドによって挿入されることがあります。

リソースがどこで実行されるかを予測できないのはなぜですか。

リソースに対して場所の制約が設定されていない場合、その配置は、(ほとんど)ランダムなノード選択によって決まります。どのノードでリソースを実行することが望ましいか、常に明示的に指定することをお勧めします。このことは、すべてのリソースに対して、場所の初期設定を行う必要があるという意味ではありません。関連する(コロケーション)リソースのセットに対して優先指定を設定すれば十分です。ノードの優先指定は次のようになります。

```
location rsc-prefers-alice rsc 100: alice
```

## A.4 STONITHとフェンシング

STONITHリソースが開始しないのはなぜですか。

開始(または有効化)操作には、デバイスのステータスのチェックが含まれます。デバイスの準備ができていない場合、STONITHリソースの開始は失敗します。

同時に、STONITHプラグインは、ホストリストを生成するように要求されます。リストが空の場合、STONITHリソースが対象にできるものがないことになるので、いずれにせよシューティングは行われません。STONITHが動作しているホストの名前は、リストから除外されます。ノードが自分自身をシューティングすることはできないからです。

停電デバイスのような、シングルホスト管理デバイスを使用する場合、フェンシングの対象とするデバイスではSTONITHリソースの動作を許可しないようにしてください。-INFINITYの、ノードの場所優先設定(制約)を使用してください。クラスタは、STONITHリソースを、起動できる別の場所に移動します。その際にはそのことが通知されます。

STONITHリソースを設定したのにフェンシングが行われしないのはなぜですか。

それぞれのSTONITHリソースは、ホストリストを持つ必要があります。このリストは、手動でSTONITHリソースの設定に挿入される場合、またはデバイス自体から取得される場合があります(たとえば出力名から)。この点は、STONITHプラグインの性質に応じて決まります。stonithdは、このリストを基に、どのSTONITHリソースがターゲットノードのフェンシングを行えるかを判断します。ノードがリストに含まれている場合に限って、STONITHリソースはノードのシューティング(フェンシング)を行います。

stonithdは、動作しているSTONITHリソースから提供されたホストリスト内にノードを見つけられなかった場合、他のノードのstonithdインスタンスに問い合わせます。他のstonithdインスタンスのホストリストにもターゲットノードが含まれていなかった場合、フェンシング要求は、開始ノードでタイムアウトのために終了します。

STONITHリソースが失敗することがあるのはなぜですか。

ブロードキャストトラフィックが多すぎると、電源管理デバイスが機能しなくなることがあります。監視操作を少なくして、余裕を持たせてください。フェンシングが一時的にのみ必要な場合(必要が生じないのが最善ですが)、デバイスのステータスは数時間に1回チェックすれば十分です。

また、この種のデバイスの中には、同時に複数の相手と通信するのを拒否するものもあります。このことは、ユーザが端末またはブラウザセッションを開いたままにしている、クラスタがステータスのテストを行おうとした場合には、問題となり得ます。

## A.5 履歴

障害の発生したリソースからステータス情報またはログを取得するにはどうしたらよいですか。

history コマンド、およびそのサブコマンド resource を使用します。

```
root # crm history resource NAME1
```

これにより、指定したリソースのみの完全な遷移ログが得られます。ただし、複数のリソースを調査することも可能です。その場合、最初のリソース名の後に目的のリソース名を追加します。一定の命名規則(を参照してください)に従っていれば、`resource` コマンドでリソースのグループを調査するのが容易になります。たとえば、次のコマンドは、`db` で始まるすべてのプリミティブを調査します。

```
root # crm history resource db*
```

`/var/cache/crm/history/live/alice/ha-log.txt` のログファイルを表示します。

履歴の出力を減らすにはどうしたらよいですか。

`history` コマンドには、次の2つのオプションがあります。

- `exclude` を使用する
- `timeframe` を使用する

`exclude` コマンドを使用すると、追加の正規表現を設定して、ログから特定のパターンを除外できます。たとえば、次のコマンドは、SSH、systemd、およびカーネルのメッセージをすべて除外します。

```
root # crm history exclude ssh|systemd|kernel.
```

`timeframe` コマンドを使用して、出力を特定の範囲に制限します。たとえば、次のコマンドは、8月23日12:00~12:30のイベントをすべて表示します。

```
root # crm history timeframe "Aug 23 12:00" "Aug 23 12:30"
```

後で検査できるように「セッション」を保存するにはどうしたらよいですか。

詳しい調査を要するバグまたはイベントが発生した場合、現在のすべての設定を保存しておく役に立ちます。このファイルをサポートに送信したり、`bzless` で表示したりできます。次に例を示します。

```
crm(live)history# timeframe "Oct 13 15:00" "Oct 13 16:00"
crm(live)history# session save tux-test
crm(live)history# session pack
Report saved in '/root/tux-test.tar.bz2'
```

## A.6 Hawk2

### 自己署名証明書の置き換え

Hawk2の最初の起動で自己署名証明書に関する警告が発行されるのを避けるには、自動生成された証明書を、独自の証明書または公式認証局(CA)によって署名された証明書で置き換えてください。

1. `/etc/hawk/hawk.key`を秘密鍵で置き換えます。
2. `/etc/hawk/hawk.pem`をHawk2が提供する証明書で置き換えます。

`root:haclient`にファイルの所有権を変更して、そのファイルがグループにアクセスできるようにします。

```
chown root:haclient /etc/hawk/hawk.key /etc/hawk/hawk.pem
chmod 640 /etc/hawk/hawk.key /etc/hawk/hawk.pem
```

## A.7 その他

すべてのクラスタノードでコマンドを実行するにはどうしたらよいですか。

この作業を実行するには、`pssh`コマンドを使用します。必要であれば、`pssh`をインストールしてください。ファイル(たとえば`hosts.txt`)を作成し、その中に操作する必要のあるノードのIPアドレスまたはホスト名を含めます。`ssh`を使用して`hosts.txt`ファイルに含まれている各ホストにログインしていることを確認します。準備ができれば、`pssh`を実行します。`hosts.txt`ファイルを(オプション `-h`で)指定し、対話モードを使用してください(オプション `-i`)。次のようになります。

```
pssh -i -h hosts.txt "ls -l /corosync/*.conf"
[1] 08:28:32 [SUCCESS] root@venus.example.com
-rw-r--r-- 1 root root 1480 Nov 14 13:37 /etc/corosync/corosync.conf
[2] 08:28:32 [SUCCESS] root@192.168.2.102
-rw-r--r-- 1 root root 1480 Nov 14 13:37 /etc/corosync/corosync.conf
```

クラスタはどのような状態でしょうか。

クラスタの現在のステータスを確認するには、`crm_mon`か`crm status`のどちらかを使用します。これによって、現在のDCと、現在のノードに認識されているすべてのノードとリソースが表示されます。

クラスタ内の一部のノードが相互に通信できないのはなぜですか。

これにはいくつかの理由が考えられます。

- まず設定ファイル `/etc/corosync/corosync.conf` を調べます。マルチキャストまたはユニキャストアドレスがクラスタ内のすべてのノードで同一かどうか確認します(キー `mcastaddr` を含む `interface` セクションを調べてください)。
- ファイアウォール設定を確認します。
- スイッチがマルチキャストまたはユニキャストアドレスをサポートしているか確認します。
- ノード間の接続が切断されていないかどうか確認します。その原因の大半は、ファイアウォールの設定が正しくないことです。また、これはスプリットブレインの理由にもなり、クラスタがパーティション化されます。

OCFS2デバイスをマウントできないのはなぜですか。

ログメッセージ(`sudo journalctl -n`)に次の行があるか確認してください。

```
Jan 12 09:58:55 alice lrmd: [3487]: info: RA output: [...]
ERROR: Could not load ocfs2_stackglue
Jan 12 16:04:22 alice modprobe: FATAL: Module ocfs2_stackglue not found.
```

この場合、カーネルモジュール `ocfs2_stackglue.ko` がありません。インストールしたカーネルに応じて、パッケージ `ocfs2-kmp-default`、`ocfs2-kmp-pae`、または `ocfs2-kmp-xen` をインストールします。

すべてのクラスタノードの分析を含むレポートを作成するにはどうしたらよいですか。

crmシェルで、`crm report` を使用してレポートを作成します。このツールは以下を収集します。

- クラスタ全体のログファイル
- パッケージ状態
- DLM/OCFS2状態
- システム情報
- CIB履歴
- コアダンプレポートの解析(debuginfoパッケージがインストールされている場合)

通常は、次のコマンドで `crm report` を実行します。

```
root # crm report -f 0:00 -n alice -n bob
```

このコマンドは、ホストaliceおよびbob上の午前0時以降のすべての情報を抽出し、現在のディレクトリに `crm_report-DATE.tar.bz2` という名前の\*.tar.bz2 アーカイブを作成します(例: `crm_report-Wed-03-Mar-2012`)。特定のタイムフレームのみを対象とする場合は、`-t` オプションを使用して終了時間を追加します。





## 警告: 機密の情報は削除してください

`crm report` ツールは、CIBと入力ファイルから機密の情報を削除しようと試みますが、完全に削除できるわけではありません。他にも機密の情報が含まれている場合には、付加的なパターンを指定してください。ログファイルと `crm_mon`、`ccm_tool`、および `crm_verify` の出力は、フィルタされません。

データをいずれの方法でも共有する前に、アーカイブをチェックして、公表したくない情報があればすべて削除してください。

さらに追加のオプションを使用して、コマンドの実行をカスタマイズします。たとえば、Pacemaker クラスタがある場合は、確実にオプション `-A` を追加する必要があるでしょう。別のユーザがクラスタに対するパーミッションを持っている場合は、(`root` および `hacluster` に加えて) `-u` オプションを使用してこのユーザを指定します。非標準のSSHポートを使用する場合は、`-X` オプションを使用して、ポートを追加します(たとえば、ポート3479では、`-X "-p 3479"` を使用)。その他のオプションは、`crm report` のマニュアルページに記載されています。

`crm report` で、関連するすべてのログファイルを分析し、ディレクトリ(またはアーカイブ)を作成したら、`ERROR` という文字列(大文字)があるかどうかログファイルをチェックします。レポートの最上位ディレクトリにある最も重要なファイルは次のとおりです。

### analysis.txt

すべてのノードで同一である必要があるファイルを比較します。

### corosync.txt

Corosync設定ファイルのコピーを格納します。

### crm\_mon.txt

`crm_mon` コマンドの出力を格納します。

### description.txt

ノード上のすべてのクラスタパッケージのバージョンを格納します。ノード固有の `sysinfo.txt` ファイルもあります。これは最上位ディレクトリにリンクしています。

このファイルは、発生した問題を説明して<https://github.com/ClusterLabs/crmsh/issues> に送信するためのテンプレートとして使用できます。

### members.txt

すべてのノードのリストです。

### sysinfo.txt

関連するすべてのパッケージ名とそのバージョンのリストが記述されています。さらに、元のRPMパッケージとは異なる設定ファイルのリストもあります。

ノード固有のファイルは、ノードの名前を持つサブディレクトリに保存されます。ここには、それぞれのノードのディレクトリ /etc のコピーが保存されます。

## A.8 その他の情報

クラスタリソースの設定、およびHigh Availabilityクラスタの管理とカスタマイズなど、Linuxの高可用性に関するその他の情報については、<http://clusterlabs.org/wiki/Documentation>  を参照してください。



## B 命名規則

このガイドでは、クラスタノードと名前、クラスタリソース、および制約に次の命名規則を使用します。

### クラスタノード

クラスタノードは名を使用します:

alice、bob、charlie、doro、およびeris

### クラスタサイトの名前

クラスタサイトには、都市の名前が付けられます:

アムステルダム、ベルリン、キャンベラ、福岡、ギザ、ハノイ、およびイスタンブール

### クラスタリソース

|             |                    |
|-------------|--------------------|
| プリミティブ      | プレフィックスなし          |
| グループ        | プレフィックス <u>g-</u>  |
| クローン        | プレフィックス <u>cl-</u> |
| マルチステートリソース | プレフィックス <u>ms-</u> |

### 制約

|           |                     |
|-----------|---------------------|
| 順序の制約     | プレフィックス <u>o-</u>   |
| 場所の制約     | プレフィックス <u>loc-</u> |
| コロケーション制約 | プレフィックス <u>col-</u> |

## C クラスタ管理ツール(コマンドライン)

High Availability Extensionには、クラスタをコマンドラインから管理する際に役立つ、包括的なツールセットが付属しています。この章では、CIBおよびクラスタリソースでのクラスタ構成を管理するために必要なツールを紹介します。リソースエージェントを管理する他のコマンドラインツールや、セットアップのデバッグ(およびトラブルシューティング)に使用するツールについては、[付録A トラブルシューティング](#)で説明されています。



### 注記: crmshの使用

これは、エキスパート専用のツールです。通常、crmシェル(crmsh)を使用したクラスタ管理が推奨されている方法です。

次のリストは、クラスタ管理に関連するいくつかの作業を示しており、これらの作業を実行するために使用するツールを簡単に説明しています。

#### クラスタのステータスの監視

`crm_mon` コマンドでは、クラスタのステータスと設定を監視できます。出力には、ノード数、uname、uuid、ステータス、クラスタで設定されたリソース、それぞれの現在のステータスが含まれます。`crm_mon` の出力は、コンソールに表示したり、HTMLファイルに出力したりできます。statusセクションのないクラスタ設定ファイルが指定された場合、`crm_mon` はファイルに指定されたノードとリソースの概要を作成します。このツールの使用法およびコマンド構文に関する詳細については、`crm_mon` マニュアルページを参照してください。

#### CIBの管理

`cibadmin` コマンドは、CIBを操作するための低レベル管理コマンドです。CIBのすべてまたは一部のダンプ、CIBのすべてまたは一部の更新、すべてまたは一部の変更、CIB全体の削除、その他のCIB管理操作に使用できます。このツールの使用法およびコマンド構文に関する詳細については、`cibadmin` マニュアルページを参照してください。

#### 設定の変更の管理

`crm_diff` コマンドは、XMLパッチの作成と適用をサポートします。クラスタの環境設定の2つのバージョンの違いを視覚的に確認する場合や、変更を保存しておき、後で `cibadmin` を使用して適用する場合には便利です。このツールの使用法およびコマンド構文に関する詳細については、`crm_diff` マニュアルページを参照してください。

## CIB属性の操作

`crm_attribute` コマンドで、CIBで使用されているノード属性およびクラスタ設定オプションを問い合わせる操作できます。このツールの使用法およびコマンド構文に関する詳細については、`crm_attribute` マニュアルページを参照してください。

## クラスタ設定の検証

`crm_verify` コマンドは、設定データベース(CIB)の整合性およびその他の問題を確認します。設定を含むファイルを確認したり、実行中のクラスタに接続したりできます。2種類の問題を報告します。エラーを解決しないと High Availability Extensionが正常に機能できず、警告の解決は管理者が担当します。`crm_verify` は新規または変更された設定の作成を支援します。実行中のクラスタのCIBのローカルコピーを作成し、編集し、`crm_verify` を使用して検証し、新規設定を `cibadmin` を使用して適用できます。このツールの使用法およびコマンド構文に関する詳細については、`crm_verify` マニュアルページを参照してください。

## リソース設定の管理

`crm_resource` コマンドは、クラスタ上でリソース関連のさまざまなアクションを実行します。設定されたリソースの定義の変更、リソースの始動と停止、リソースの削除およびノード間でのマイグレートを実行できます。このツールの使用法およびコマンド構文に関する詳細については、`crm_resource` マニュアルページを参照してください。

## リソースの失敗回数の管理

`crm_failcount` コマンドは、所定のノードのリソースごとの失敗回数を問い合わせます。このツールは、失敗回数のリセットにも使用でき、リソースが頻繁に失敗したノード上で再度実行できるようにします。このツールの使用法およびコマンド構文に関する詳細については、`crm_failcount` マニュアルページを参照してください。

## ノードのスタンバイステータスの管理

`crm_standby` コマンドは、ノードのスタンバイ属性を操作します。スタンバイモードのノードはすべて、リソースをホストすることができず、そのノードにあるリソースは削除する必要があります。スタンバイモードはカーネルの更新などの保守作業を行う場合に便利です。ノードを再びクラスタの完全にアクティブなメンバーにするには、ノードからスタンバイ属性を削除します。このツールの使用法およびコマンド構文に関する詳細については、`crm_standby` マニュアルページを参照してください。

## D rootアクセスなしでのクラスタレポートの実行

すべてのクラスタノードはSSHによって互いにアクセスできる必要があります。`crm report` (トラブルシューティング用)などのツールおよびHawk2の[履歴エクスプローラ]は、ノード間でパスワード不要のSSHアクセスを必要とします。それがない場合、現在のノードからしかデータを収集できません。

パスワード不要のSSH `root` アクセスが規定要件を順守しない場合は、次善策を使用してクラスタレポートを実行できます。これは次の基本手順で構成されます。

1. 専用のローカルユーザアカウント(`crm report` 実行用)を作成する。
2. できれば標準以外のSSHポートを使用して、そのユーザアカウントにパスワード不要のSSHアクセスを設定する。
3. そのユーザに `sudo` を設定する。
4. そのユーザとして `crm report` を実行する。

デフォルトでは、`crm report` を実行すると、まず `root` としてリモートノードへのログインを試行し、続いてユーザ `hacluster` としてログインを試行します。ただし、ローカルセキュリティポリシーによってSSHを使用した `root` ログインが禁止されている場合、スクリプトの実行はすべてのリモートノードで失敗します。ユーザ `hacluster` としてスクリプトを実行しようとしても失敗します。これはサービスアカウントであり、そのシェルはログインが禁止されている `/bin/false` に設定されているためです。High Availabilityクラスタのすべてのノードで `crm report` スクリプトを正常に実行する唯一のオプションは、専用のローカルユーザを作成することだけです。

### D.1 ローカルユーザアカウントの作成

次の例では、コマンドラインから `hareport` という名前のローカルユーザを作成します。パスワードは、セキュリティ要件を満たしていれば何でも構いません。または、YaSTでユーザアカウントを作成してパスワードを設定することもできます。

手順 D.1: クラスタレポート実行用の専用ローカルユーザアカウントの作成

1. シェルを起動し、ホームディレクトリ `/home/hareport` を持つユーザ `hareport` を作成します。

```
root # useradd -m -d /home/hareport -c "HA Report" hareport
```

2. ユーザのパスワードを設定します。

```
root # passwd hareport
```

3. プロンプトが表示されたら、ユーザのパスワードを入力し、確認のために再度入力します。

## ！ 重要: 各クラスターノードで同じユーザが必要

すべてのノードで同じユーザアカウントを作成するため、各ノードで上の手順を繰り返してください。

## D.2 パスワード不要のSSHアカウントの設定

手順 D.2: SSHデーモンに対する非標準ポートの設定

デフォルトでは、SSHデーモンおよびSSHクライアントはポート 22 で通信およびリスンします。ネットワークセキュリティガイドラインによって、デフォルトのSSHポートを大きい番号の別のポートに変更することが要求されている場合は、デーモンの設定ファイル /etc/ssh/sshd\_config を変更する必要があります。

1. デフォルトポートを変更するには、ファイルで Port 行を検索してコメント解除し、目的に応じて編集します。たとえば、次のように設定します。

```
Port 5022
```

2. 組織において root ユーザが他のサーバにアクセスすることが許可されていない場合、このファイルで PermitRootLogin エントリを検索してコメント解除し、no に設定します。

```
PermitRootLogin no
```

3. または、次のコマンドを実行して、各行をファイルの末尾に追加します。

```
root # echo "PermitRootLogin no" >> /etc/ssh/sshd_config
root # echo "Port 5022" >> /etc/ssh/sshd_config
```

4. /etc/ssh/sshd\_config を変更したら、SSHデーモンを再起動して新しい設定を有効にします。

```
root # systemctl restart sshd
```

## ！ 重要: 各クラスターノードで同じ設定が必要

各クラスターノードで、このSSHデーモンの設定を繰り返してください。

#### 手順 D.3: SSHクライアントに対する非標準ポートの設定

クラスタ内のすべてのノードでSSHポートを変更する場合、SSH設定ファイル `/etc/ssh/sshd_config` を変更するのが便利です。

1. デフォルトポートを変更するには、ファイルで `Port` 行を検索してコメント解除し、目的に応じて編集します。たとえば、次のように設定します。

```
Port 5022
```

2. または、次のコマンドを実行して、各行をファイルの末尾に追加します。

```
root # echo "Port 5022" >> /etc/ssh/sshd_config
```



### 注記: 1つのノードでのみ設定が必要

このSSHクライアントの設定は、クラスタレポートを実行するノードでのみ必要です。

または、`-X` オプションを使用してカスタムSSHポートで `crm report` を実行することも、`crm report` がデフォルトでカスタムSSHポートを使用するように設定することもできます。詳細については、[手順D.5「カスタムSSHポートを使用したクラスタレポートの生成」](#)を参照してください。

#### 手順 D.4: 共有SSH鍵の作成

パスワードを要求されずに、SSHを使用して他のサーバにアクセスできます。これは一見、安全ではないように見えますが、ユーザは自身の公開鍵が共有されているサーバにしかアクセスできないため、実際は非常に安全なアクセス方法です。共有鍵は、その鍵を使用するユーザとして作成する必要があります。

1. クラスタレポートの実行用に作成したユーザアカウントでノードの1つにログインします(上の例では、このユーザアカウントは `hareport` です)。
2. 新しい鍵を生成します。

```
hareport > ssh-keygen -t rsa
```

このコマンドは、デフォルトで2,048ビットの鍵を生成します。鍵のデフォルトの場所は `~/.ssh/` です。鍵にパスフレーズを設定するよう求められます。ただし、パスワードなしでログインする場合は、鍵にパスフレーズがあってはならないため、パスフレーズを入力しないでください。

3. 鍵が生成されたら、公開鍵を他の「それぞれの」ノードにコピーします(鍵を作成したノードを「含む」)。

```
hareport > ssh-copy-id -i ~/.ssh/id_rsa.pub HOSTNAME_OR_IP
```

このコマンドでは、各サーバのDNS名、エイリアス、またはIPアドレスを使用できます。コピー中に、各ノードのホスト鍵を受諾するよう求められるので、`hareport` ユーザアカウントのパスワードを入力する必要があります(パスワードを入力する必要があるのは、ここだけです)。

4. 鍵をすべてのクラスタノードと共有したら、パスワード不要のSSHを使用して、ユーザ `hareport` として他のノードにログインできるかどうかをテストします。

```
hareport > ssh HOSTNAME_OR_IP
```

証明書の受諾やパスワードの入力を要求されることなく、自動的にリモートサーバに接続されます。



### 注記: 1つのノードでのみ設定が必要

毎回同じノードからクラスタレポートを実行する場合は、上の手順は、このノードでのみ実行すれば十分です。そうでない場合は、各ノードでこの手順を繰り返します。

## D.3 `sudo`の設定

`sudo` コマンドは、通常のユーザを素早く `root` にしてコマンドを発行できるようにします。パスワードの入力は、必要な場合と不要な場合があります。すべてのルートレベルのコマンドに `sudo` アクセスを付与することも、特定のコマンドにのみ付与することもできます。一般的には、`sudo` はエイリアスを使用してコマンド文字列全体を定義します。

`sudo` を設定するには、`visudo` (viでは「ありません」)またはYaSTを使用します。



### 警告: viは使用しない

コマンドラインから `sudo` を設定するには、`visudo` を使用して、`root` として `sudoers` ファイルを編集する必要があります。他のエディタを使用すると、構文エラーやファイルパーミッションエラーが発生し、`sudo` を実行できないことがあります。

1. `root` としてログインします。
2. `/etc/sudoers` ファイルを開くため、「`visudo`」と入力します。
3. カテゴリ `Host alias specification`、`User alias specification`、`Cmnd alias specification`、および `Runas alias specification` を探します。



4. 次のエントリを `/etc/sudoers` 内のそれぞれのカテゴリに追加します。

```
Host_Alias CLUSTER = alice,bob,charlie ❶
User_Alias HA = hareport ❷
Cmnd_Alias HA_ALLOWED = /bin/su, /usr/sbin/crm report * ❸
Runas_Alias R = root ❹
```

- ❶ ホストエイリアスは、sudoユーザがコマンド発行権利を持つサーバ(またはサーバの範囲)を定義します。ホストエイリアスでは、DNS名またはIPアドレスを使用するか、ネットワーク範囲全体を指定できます(例: `172.17.12.0/24`)。アクセスの範囲を制限するには、クラスターノードのホスト名のみを指定する必要があります。
- ❷ ユーザエイリアスでは、複数のローカルユーザアカウントを1つのエイリアスに追加できます。ただし、この場合、使用するアカウントは1つだけであるため、エイリアスの作成を避けることができます。上の例では、クラスタレポート実行用に作成した `hareport` ユーザを追加しています。
- ❸ コマンドエイリアスは、ユーザが実行できるコマンドを定義します。これは、非ルートユーザが `sudo` を使用する際にアクセスできるコマンドを制限する場合に便利です。この場合、`hareport` ユーザアカウントには、コマンド `crm report` および `su` に対するアクセスが必要です。
- ❹ `runas` エイリアスは、コマンドの実行に使用するアカウントを指定します。この場合は、`root` です。

5. 次の2行を検索します。

```
Defaults targetpw
ALL ALL=(ALL) ALL
```

これらは、作成したい設定と衝突するため、無効にします。

```
#Defaults targetpw
#ALL ALL=(ALL) ALL
```

6. `User privilege specification` を探します。

7. 上のエイリアスを定義したら、そこに次のルールを追加できます。

```
HA CLUSTER = (R) NOPASSWD:HA_ALLOWED
```

`NOPASSWD` オプションは、ユーザ `hareport` がパスワードを入力せずにクラスタレポートを実行できるようにします。



## ！ 重要: 各クラスターノードで同じsudo設定が必要

クラスター内のすべてのノードでこのsudo設定を行う必要があります。sudoに他の変更は必要なく、再起動が必要なサービスもありません。

## D.4 クラスタレポートの生成

上で行った設定でクラスタレポートを実行するには、ノードの1つにユーザ `hareport` としてログインする必要があります。クラスタレポートを起動するには、`crm report` コマンドを使用します。次に例を示します。

```
root # crm report -f 0:00 -n "alice bob charlie"
```

このコマンドは、指定したノード上の 午前0時 以降の情報をすべて抽出し、現在のディレクトリに `pcmk-DATE.tar.bz2` という名前の `*.tar.bz2` アーカイブを作成します。

手順 D.5: カスタムSSHポートを使用したクラスタレポートの生成

1. カスタムSSHポートを使用する場合、`crm report` で `-X` を使用して、クライアントのSSHポートを変更します。たとえば、カスタムSSHポートが `5022` の場合、次のコマンドを使用します。

```
root # crm report -X "-p 5022" [...]
```

2. `crm report` のカスタムSSHポートを永続的に設定するには、`crm` 対話型シェルを開始します。

```
crm options
```

3. 次のように入力します。

```
crm(live)options# set core.report_tool_options "-X -oPort=5022"
```

## 用語集

### AutoYaST

AutoYaSTは、ユーザの介入なしで、1つ以上のSUSE Linux Enterpriseシステムを自動的にインストールするためのシステムです。

### bindnetaddr (バインドネットワークアドレス)

Corosyncエグゼクティブのバインド先のネットワークアドレス。

### boothd (ブースデーモン)

Geoクラスタ内のそれぞれの参加クラスタおよびアービトラータが、サービスである **boothd** を実行します。これは、別のサイトで実行しているブースデーモンに接続し、接続性の詳細を交換します。

### CCM (コンセンサスクラスタメンバーシップ)

CCMは、どのノードがクラスタを設定するか決定し、この情報をクラスタで共有します。ノードまたはクォーラムの新規追加および損失は、CCMによって通知されます。CCMモジュールはクラスタの各ノード上で実行されます。

### CIB (クラスタ情報ベース)

クラスタの設定とステータス(クラスタオプション、ノード、リソース、制約、相互の関係性)の全体的な表現。XMLで作成され、メモリに常駐します。マスタCIBは、**DC (指定コーディネータ)**で保持および保守され、他のノードに複製されます。CIBに対する通常の読み書き操作は、マスタCIBによってシリアルに処理されます。

### conntrackツール

カーネル内の接続トラッキングシステムとやり取りできるようにして、iptablesでのステートフルなパケット検査を可能にします。High Availability Extensionによって、クラスタノード間の接続ステータスを同期化するために使用されます。

### CRM (クラスタリソースマネージャ)

すべての非ローカルインタラクションの調整に責任を負う主要管理エンティティ。High Availability Extensionでは、PacemakerをCRMとして使用します。クラスタの各ノードにはノード独自のCRMインスタンスがありますが、DC上で実行されるCRMインスタンスは、決定を他の非ローカルCRMに中継し、それらからの入力を処理するために選択されたものです。CRMは、いくつかのコンポーネント(CRM自身のノードとその他のノード両方のローカルリソースマネージャ、非ローカルCRM、管理コマンド、フェンシング機能、メンバーシップ層、およびブース)と対話します。

### crmd (クラスタリソースマネージャデーモン)

CRMは、crmdというデーモンとして実装されます。crmdは各クラスタノード上にインスタンスを持ちます。マスタとして動作するcrmdインスタンスを1つ選択することにより、クラスタのすべての意思

決定が一元化されます。選択したcrmdプロセス(またはそのプロセスが実行されているノード)で障害が発生したら、新しいcrmdプロセスが確立されます。

#### crmsh

コマンドラインユーティリティcrmshは、クラスタ、ノード、およびリソースを管理します。

詳細については、[第8章「クラスタリソースの設定と管理\(コマンドライン\)」](#)を参照してください。

#### Csync2

クラスタ内のすべてのノード、およびGeoクラスタ全体に設定ファイルを複製するために使用できる同期ツールです。

#### DC (指定コーディネータ)

クラスタ内のCRMは、指定コーディネータ(DC)として選択されます。DCは、ノードのフェンシングやリソースの移動など、クラスタ全体におよぶ変更が必要かどうかを判断できる、クラスタ内で唯一のエンティティです。DCは、CIBのマスターコピーが保持されるノードでもあります。その他すべてのノードは、現在のDCから設定とリソース割り当て情報を取得します。DCは、メンバーシップの変更後、クラスタ内のすべてのノードから選抜されます。

#### DLM (分散ロックマネージャ)

DLMは、クラスタファイルシステムのディスクアクセスを調整し、ファイルロックを管理して、パフォーマンスと可用性を向上します。

#### DRBD

DRBD®は、高可用性クラスタを構築するためのブロックデバイスです。ブロックデバイス全体が専用ネットワーク経由でミラーリングされ、ネットワークRAID-1として認識されます。

#### Geoクラスタ

それぞれにローカルクラスタを持つ、複数の地理的に離れたサイトで構成されます。サイトはIPによって交信します。サイト全体のフェールオーバーはより高いレベルのエンティティ(ブース)によって調整されます。Geoクラスタは限られたネットワーク帯域幅および高レイテンシに対応する必要があります。ストレージは同期的にレプリケートされます。

#### geoクラスタ(地理的に離れたクラスタ)

詳細については、[Geoクラスタ](#)を参照してください。

#### LRM (ローカルリソースマネージャ)

リソースに対する操作の実行を担当します。リソースエージェントスクリプトを使用してこれらの操作を実行します。LRMはポリシーを認識していないという点で、「ダム」です。何をすべきか認識させるにはDCが必要です。

#### mcastaddr (マルチキャストアドレス)

Corosyncエグゼクティブによるマルチキャストに使用されるIPアドレス。このIPアドレスはIPv4またはIPv6のいずれかに設定できます。

#### mcastport (マルチキャストポート)

クラスタ通信に使用されるポート。

#### multicast (マルチキャスト)

ネットワーク内で一対多数の通信に使用される技術で、クラスタ通信に使用できます。Corosyncはマルチキャストとユニキャストの両方をサポートしています。

#### PE (ポリシーエンジン)

ポリシーエンジンはCIBでのポリシー変更を実装するために必要な処理を計算します。PEは(リソース)アクションのリストと、次のクラスタ状態に移るために必要な依存性を含む遷移グラフも作成します。PEは常にDC上で実行されます。

#### RA (リソースエージェント)

プロキシとして機能してリソースを管理する(リソースの開始、停止、監視などを行う)スクリプト。High Availability Extensionは、OCF (Open Cluster Framework)、LSB (Linux Standard Base initスクリプト)、およびHeartbeatという3種類のリソースエージェントをサポートしています。詳細については、[6.3.2項「サポートされるリソースエージェントクラス」](#)を参照してください。

#### Rear (Relaxおよび回復)

障害復旧イメージを作成するための管理ツールセット。

#### RRP (冗長リングプロトコル)

ネットワーク障害の一部または全体に対する災害耐性のために、複数の冗長ローカルエリアネットワークが使用できるようになります。この方法では、ひとつのネットワークが作動中である限り、クラスタ通信を維持できます。Corosyncはトータム冗長リングプロトコルをサポートします。

#### SBD (STONITHブロックデバイス)

共有ブロックストレージ(SAN、iSCSI、FCoEなど)を介したメッセージの交換を通じて、ノードフェンシングメカニズムを提供します。ディスクレスモードでも使用できます。動作異常のノードが本当に停止したかどうかを確認するために、各ノードではハードウェアまたはソフトウェアのウォッチドッグが必要です。

#### SFEX (共有ディスクファイル排他制御)

SFEXはSANにおけるストレージ保護機能を提供します。

#### SPOF (シングルポイント障害)

失敗するとクラスタ全体の障害をトリガしてしまう、クラスタのコンポーネント。

## STONITH

「Shoot the other node in the head」の略です。動作異常のノードをシャットダウンすることでクラスタに問題を発生させないようにするフェンシングメカニズムを指しています。

## アクティブ/アクティブ、アクティブ/パッシブ

サービスがノード上で実行される方法の概念。アクティブ/パッシブシナリオでは、1つ以上のサービスがアクティブノード上で実行され、パッシブノードはアクティブノードの失敗を待機します。アクティブ/アクティブでは、各ノードがアクティブであると同時にパッシブです。たとえば、一部のサービスは実行されていますが、それ以外のサービスは他のノードから引き継ぐことができます。DRBDのプライマリ/セカンダリとデュアルプライマリに類似しています。

## アービトラータ

Geoクラスタ内の追加インスタンスで、サイト間にまたがるリソースのフェールオーバーなどの決定に関する合意の形成を手助けします。アービトラータは専用モードで1つまたは複数のブースインスタンスを実行する単一のマシンです。

## クォーラム

クラスタでは、クラスタパーティションは、ノード(投票)の過半数を保有する場合、クォーラムを持つ(「定足数に達している」と定義されます。クォーラムはただ1つのパーティションで識別されます。複数の切断されたパーティションまたはノードが処理を続行してデータおよびサービスが破損されないようにする、アルゴリズムの一部です(スプリットブレイン)。クォーラムはフェンシングの前提条件で、このためクォーラムは一意になります。

## クラスタ

「ハイパフォーマンス」クラスタは、結果を早く出すためにアプリケーション負荷を共有するコンピュータ(実際または仮想のコンピュータ)のグループです。高可用性クラスタは、サービスの可用性を最大にすることを第一に設計されています。

## クラスタパーティション

1つ以上のノードとその他のクラスタ間で通信が失敗した場合は、常にクラスタパーティションが発生します。クラスタのノードはパーティションに分割されますが、アクティブなままです。これらは同じパーティション内のノードのみと通信可能で、切り離されたノードは認識しません。つまり、他のパーティションのノードの損失は確認できないため、スプリットブレインシナリオが作成されます([スプリットブレイン](#)も参照)。

## スイッチオーバー

クラスタ内の他のノードへの、予定されたオンデマンドのサービス移動。詳細については、[フェールオーバー](#)を参照してください。

## スプリットブレイン

クラスタノードが(ソフトウェアまたはハードウェア障害によって)互いに認識しない2つ以上のグループに分割される場合のシナリオです。STONITHによって、スプリットブレインがクラスタ全体に悪影響をおよぼさなくなります。「パーティションされたクラスタ」シナリオとも呼ばれます。

スプリットブレインという用語は、DRBDでも使用されますが、2つのノードに異なるデータが含まれることを意味します。

## チケット

Geoクラスタで使用されるコンポーネント。チケットは指定のクラスタサイトの特定のリソースを実行する権利を付与します。チケットは1度に1つのサイトだけが所有できます。リソースを特定のチケットに依存させることができます。定義されたチケットがサイトで使用できる場合のみそれぞれのリソースが始動します。その逆に、チケットが削除されると、そのチケットに依存するリソースが自動的に停止します。

## ノード

クラスタのメンバで、ユーザには見えない(実際または仮想の)コンピュータ。

## ハートビート

Corosyncの代替機能である、バージョン3のCCM。3つ以上の通信パスはサポートするが、クラスタファイルシステムはサポートしません。

## フェンシング

孤立または失敗したクラスタメンバーによる共有リソースへのアクセス防止の概念を示しています。クラスタノードが失敗した場合は、シャットダウンまたはリセットすることで問題の発生を防止します。つまり、ステータスが不明なノードからリソースがロックアウトされます。

## フェールオーバー

リソースまたはノードが1台のマシンで失敗し、影響を受けるリソースが別のノードで起動されたときに発生します。

## フェールオーバードメイン

ノード障害の発生時にクラスタサービスを実行することができる、指定されたクラスタノードのサブセット。

## ブース

Geoクラスタのサイト間のフェールオーバープロセスを管理するインスタンス。その目的は、1つのサイトのみでマルチサイトリソースをアクティブにすることです。これは、サイトをダウンする必要がある場合、クラスタサイト間のフェールオーバードメインとして処理される、いわゆるチケットを使用することで可能になります。

## メトロクラスタ

すべてのサイトがファイバチャネルで接続された、複数の建物またはデータセンターにわたってストレッチできる単一のクラスタ。ネットワークの遅延時間は通常は短くなります(約20マイルの距離で<5ms)。ストレージは頻繁にレプリケートされます(ミラーリングまたは同期レプリケーション)

## ユニキャスト

ひとつのあたえ先ネットワークにメッセージを送信する技術Corosyncはマルチキャストとユニキャストの両方をサポートしています。Corosyncでは、ユニキャストはUDP-unicast (UDPU)として実装されます。

## リソース

Pacemakerに認識されている、任意のタイプのサービスまたはアプリケーション。IPアドレス、ファイルシステム、データベースなどです。

「リソース」という用語は、DRBDでも使用されており、レプリケーション用の一般的な接続を使用しているブロックデバイスのセットの名前を表します。

## ローカルクラスタ

1つのロケーション内の単一のクラスタ(たとえば、すべてのノードが1つのデータセンターにある)。ネットワークの遅延時間は無視できます。ストレージは通常、すべてのノードに同時にアクセスされます。

## 同時実行違反

クラスタ内の1つのノードだけで実行する必要があるリソースが、複数のノード上で実行されています。

## 既存のクラスタ

「既存のクラスタ」という用語は、1つ以上のノードで構成されるクラスタを指すものとして使用されます。既存のクラスタは、通信チャネルを定義する基本的なCorosync設定を持ちますが、必ずしもリソース設定を持つとは限りません。

## 負荷分散

複数のサーバを同じサービスに参加させて、同じ作業を行わせる機能。

## 障害

自然、人、ハードウェアのエラー、ソフトウェアのバグなどによって引き起こされる重要なインフラストラクチャの想定外の障害

## 障害復旧

障害復旧は、障害発生後、ビジネス機能を通常どおりの、安定した状態に修復するプロセスです。

## 障害復旧プラン

ITインフラストラクチャへの影響を最小限に抑えながら障害から復旧する戦略。



# E GNU Licenses

## 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 noncommercially. 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 in order 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 noncommercially, 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 <http://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.