FUJITSU

FUJITSU Software PRIMECLUSTER

コンセプトガイド 4.5

Oracle Solaris / Linux

J2S2-1686-02Z0(02)
2018年8月

はじめに

本書では、PRIMECLUSTERの各種製品の概要について説明します。PRIMECLUSTER製品は、高可用性 (HA) と拡張性を保証する、オペレーティングシステムおよびハードウェアプラットフォーム非依存のクラスタリングソリューションです。PRIMECLUSTERのモジュール化されたソフトウェアアーキテクチャにより、クラスタ内の全てのコンピュータ (ノード) に導入される基本的なモジュールセット、および特定のアプリケーションをサポートするオプションモジュールで構成されます。モジュール化されたアーキテクチャにより多様なユーザに柔軟なクラスタリングソリューションを提供することが可能になります。このソリューションは現在および将来のあらゆるプラットフォームに適応可能です。

注意

本書では、PRIMECLUSTERの全てのコンポーネントについて説明します。ただし、リリースによっては利用できないコンポーネントもあります。それぞれのプラットフォームで使用できる機能については、“PRIMECLUSTER ソフトウェア説明書 / インストールガイド” を確認してください。

本書の読者

本書はエンドユーザ、システム管理者を対象にしています。本書はクラスタソリューションの基本概念の説明を目的としたもので、管理方法、構成設定、およびインストールの説明書ではありません (詳細については、“関連マニュアル”を参照してください)。

本書について

本書の構成は以下のとおりです。

章タイトル	内容
第1章 クラスタリングテクノロジーの概要	PRIMECLUSTERの主なコンポーネントなどのクラスタリングの概念とメリットについて説明します。
第2章 PRIMECLUSTERのアーキテクチャ	PRIMECLUSTERのアーキテクチャおよび主な機能について説明します。
第3章 クラスタインタコネクトの詳細	クラスタインタコネクトの概要、要件、設計上の検討事項について説明します。
第4章 RMS (Reliant Monitor Services)	RMSの基本的な概念、コンポーネント、メリットについて説明します。
第5章 RMSウィザード	RMSウィザードを構成するRMS Wizard Toolsについて説明します。
付録A リリース情報	マニュアルの変更について説明します。

関連マニュアル

以下のマニュアルには、PRIMECLUSTERに関する情報が記載されています。必要に応じて参照してください。

- PRIMECLUSTER 導入運用手引書
- PRIMECLUSTER 導入運用手引書＜FUJITSU Cloud Service K5環境編＞
- PRIMECLUSTER Web-Based Admin View 操作手引書
- PRIMECLUSTER Cluster Foundation 導入運用手引書
- PRIMECLUSTER RMS 導入運用手引書
- PRIMECLUSTER Global Disk Services 説明書
- PRIMECLUSTER Global File Services 説明書
- PRIMECLUSTER Global Link Services 説明書 (伝送路二重化機能編)
- PRIMECLUSTER Global Link Services 説明書 (伝送路二重化機能 仮想NIC方式編)
- PRIMECLUSTER Global Link Services 説明書 (マルチパス機能編)
- PRIMECLUSTER DR/PCI Hot Plug ユーザーズガイド

- PRIMECLUSTER 活用ガイド＜メッセージ集＞
- PRIMECLUSTER 活用ガイド＜コマンドリファレンス編＞
- PRIMECLUSTER 活用ガイド＜トラブルシューティング編＞

注意

PRIMECLUSTERの関連ドキュメントには上記マニュアル以外に以下のドキュメントがあります。

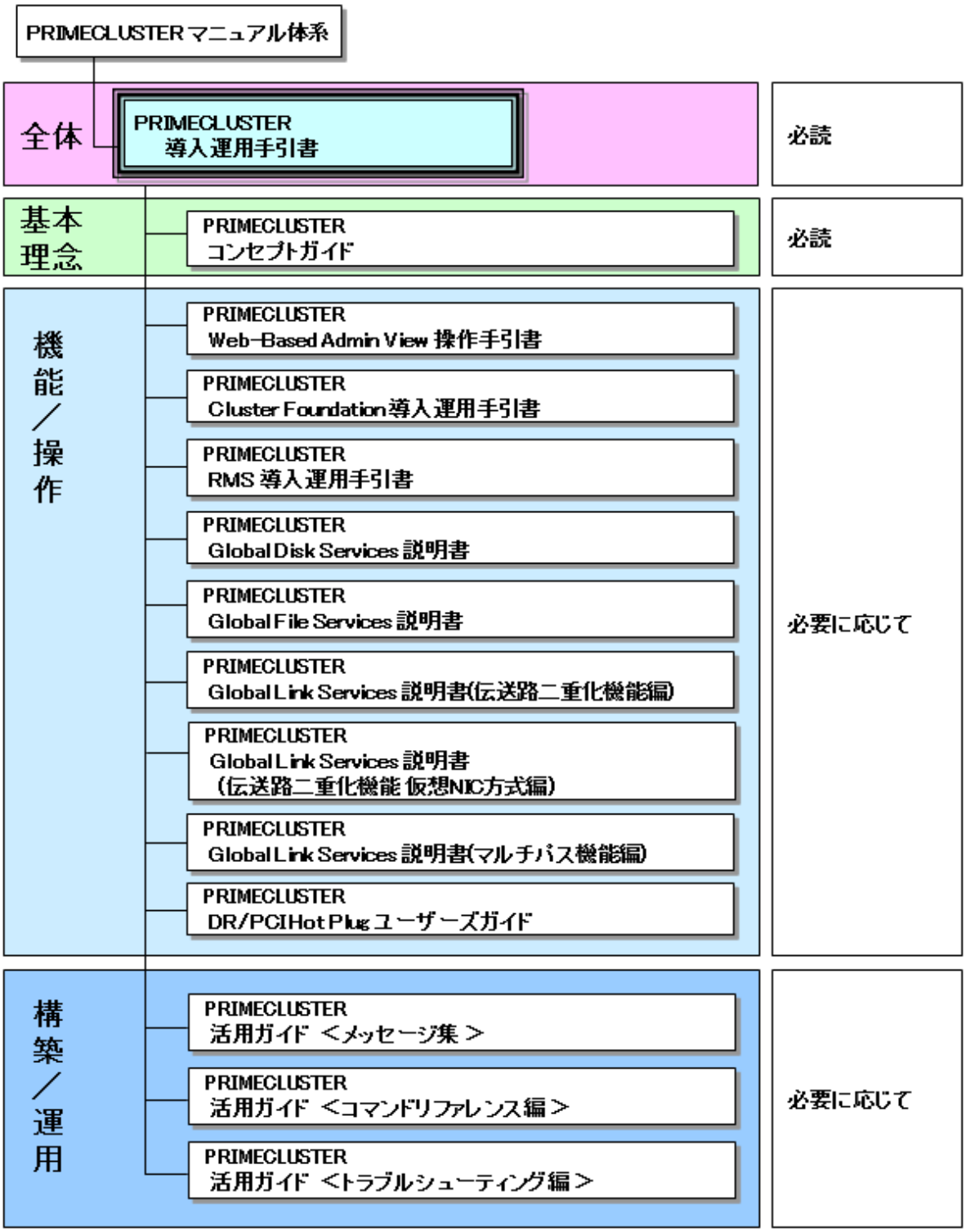
- PRIMECLUSTER ソフトウェア説明書/インストールガイド

PRIMECLUSTERの各製品に添付されるソフトウェア説明書およびインストールガイドです。

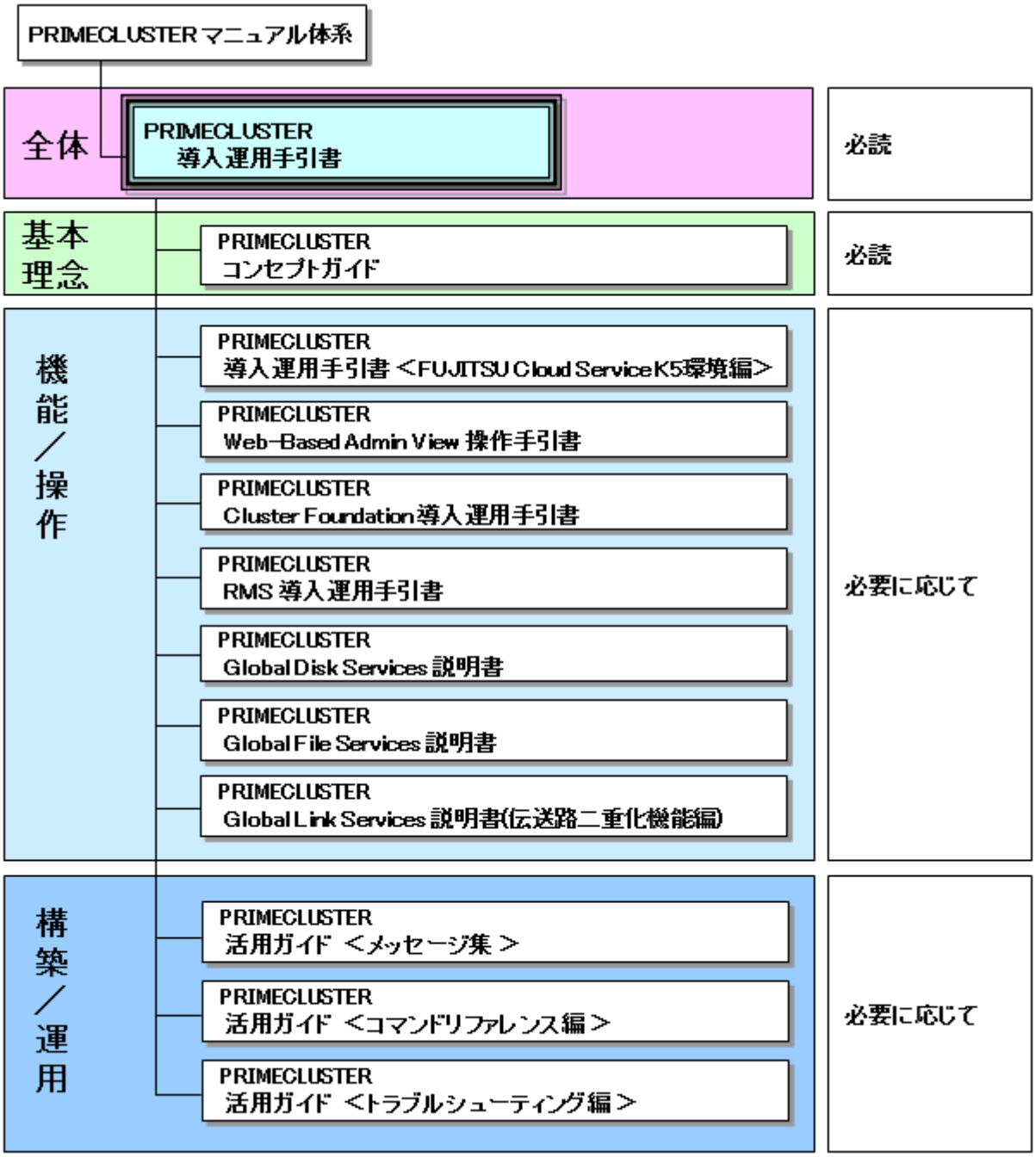
データは各製品の“DVD”に格納されています。また、ファイル名については、「製品のご案内」を参照ください。

マニュアルの体系

Solaris版



Linux(RHEL)版



マニュアルの印刷について

マニュアルの印刷をする場合には、PRIMECLUSTER製品用DVDの中に入っているPDFファイルを利用してください。PDFファイル名とマニュアルとの関係については、製品に添付されているPRIMECLUSTERのソフトウェア説明書を参照してください。

PDFファイルの参照・印刷には、Adobe Readerが必要です。Adobe Readerの入手方法については、Adobe Systems Incorporated. (アドビシステムズ社)のホームページを参照してください。

本書の表記について

表記

プロンプト

実行にシステム管理者 (ルート) 権限が必要なコマンドライン例の場合、先頭にシステム管理者プロンプトを示すハッシュ記号 (#) が付いています。いくつかの例で、node#という表記は、指定されたノードのrootプロンプトを表しています。たとえば、コマンド名の前にfuji2#が記述されていると、そのコマンドがfuji2という名前のノード上で、rootユーザとして実行されたことを示しています。システム管理者権限を必要としないエントリの場合、先頭にドル記号 ($) が付いています。

マニュアルページのセクション番号

オペレーティングシステムコマンドの後ろにマニュアルページのセクション番号が括弧付きで示されています。
例: cp(1)

キーボード

印字されない文字のキーストロークは＜Enter＞や＜F1＞などのキーアイコンで表示されます。たとえば、＜Enter＞は*Enter*というラベルの付いたキーを押すことを意味し、＜Ctrl＞+＜B＞は*Ctrl*または*Control*というラベルの付いたキーを押しながら＜B＞キーを押すことを意味します。

書体/記号

以下の書体は特定要素の強調に使用されます。

書体 / 記号	使用方法
均等幅	コンピュータ出力、およびプログラムリスト: テキスト本文中のコマンド、ファイル名、マニュアルページ名、他のリテラルプログラミング項目
斜体, <斜体>	具体的な数値/文字列に置き換える必要のある変数 —入力値—
<均等幅>	具体的な数値/文字列に置き換える必要のある変数 —表示値—
太字	記述どおりに入力する必要のあるコマンドライン項目
“均等幅”	参照先のタイトル名、マニュアル名、画面名等
[均等幅]	ツールバー名、メニュー名、コマンド名、アイコン名
＜均等幅＞	ボタン名

例1

以下に/etc/passwdファイルのエントリの一部を示します。

```
root:x:0:1:0000-Admin(0000):/:/sbin/ksh
sysadm:x:0:0:System Admin.:/usr/admin:/usr/sbin/sysadm
setup:x:0:0:System Setup:/usr/admin:/usr/sbin/setup
daemon:x:1:1:0000-Admin(0000):/:
```

例2

cat(1) コマンドでファイルの内容を表示するには、以下のコマンドラインを入力します。

```
$ cat ＜ファイル名＞
```

コマンド構文

コマンド構文には以下の規則があります。

記号	名前	意味
[ ]	角括弧	オプション項目を囲む。
{ }	波括弧	択一選択の複数選択肢を囲む。各項目は縦線 (|) で区切られる。
|	縦線	波括弧で囲まれている場合は、択一選択の各選択肢の区切り。波括弧で囲まれていない場合は、1つのプログラムの出力が他のプログラムの入力にパイプされることを示すリテラル要素。

記号	名前	意味
( )	丸括弧	繰り返しの際にグループ化される項目を囲む。
...	省略符号	項目の繰り返しを示す。1グループの項目を繰り返す場合には、項目グループを丸括弧で囲む。

記号

特に注意すべき事項の前には以下の記号が付いています。

ポイントとなる内容について説明します。

注意する項目について説明します。

例題を用いて説明します。

参考となる内容を説明します。

参照するマニュアル名などを説明します。

略称

Oracle Solarisは、Solaris、Solaris Operating System、またはSolaris OSと記載することがあります。

参照するOracle Solaris (以降、Solaris) のマニュアル名称で“Solaris X” と書かれている部分は、Oracle Solaris 10 (以降、Solaris 10) 、またはOracle Solaris 11 (以降、Solaris 11) と読み替えてマニュアルを参照してください。

Red Hat Enterprise Linux を RHEL と略しています。

RHELを Linux と表記しています。

Red Hat OpenStack Platform を RHOSP と略しています。

PRIMEQUEST 3000/2000シリーズを PRIMEQUEST と略しています。

FUJITSU Cloud Service K5をK5と略しています。

輸出管理規制について

本ドキュメントを輸出または第三者へ提供する場合は、お客様が居住する国および米国輸出管理関連法規等の規制をご確認のうえ、必要な手続きをおとりください。

商標について

UNIX は、米国およびその他の国におけるオープン・グループの登録商標です。

Red Hat は米国およびそのほかの国において登録されたRed Hat, Inc. の商標です。

出版年月および版数

2017年 4月 初版
2017年12月 第2版
2018年 5月 第2.1版
2018年 8月 第2.2版

著作権表示

変更履歴

追加・変更内容	変更箇所	版数
RHOSP環境の場合の説明を追加しました。	1.1 概要 1.6 仮想化対応 1.7.1 Linux 2.3.5 PRIMECLUSTER SF	第2.1版
以下のシャットダウンエージェントを追加しました。 ・OpenStack API(SA_vmosr)	2.3.5 PRIMECLUSTER SF	第2.2版

目　次

第1章 クラスタリングテクノロジーの概要 1
1.1 概要 1
1.2 高可用性 (HA) 2
1.2.1 クラスタインタコネクト 2
1.2.2 HAマネージャ 2
1.2.2.1 データ整合性の保証 2
1.2.2.2 ウィザード 5
1.2.3 パトロール診断機能 (Solaris) 6
1.3 拡張性 (スケーラビリティ) 6
1.4 シングルノードクラスタ 6
1.5 ノード間のデータ引継ぎ 6
1.6 仮想化対応 7
1.7 異常発生時の切替え動作 11
1.7.1 Linux 11
1.7.2 Oracle Solaris 14
1.7.2.1 Oracle Solaris(物理、Oracle VM Server for SPARC環境) 14
1.7.2.2 Oracle Solaris(Oracle Solaris カーネルゾーン環境) 16
1.7.2.3 Oracle Solaris(Oracle Solaris ノングローバルゾーン環境) 19

第2章 PRIMECLUSTERのアーキテクチャ 21
2.1 アーキテクチャの概要 21
2.2 PRIMECLUSTER設計理念 22
2.2.1 モジュール方式 22
2.2.2 プラットフォーム非依存性 22
2.2.3 拡張性 (スケーラビリティ) 23
2.2.4 可用性 23
2.2.5 データの整合性保証 23
2.3 PRIMECLUSTERのモジュール 23
2.3.1 CF 24
2.3.2 Cluster Admin 24
2.3.3 Web-Based Admin View 24
2.3.4 クラスタリソース管理機構(Cluster Resource Management)(CRM) 24
2.3.5 PRIMECLUSTER SF 25
2.3.6 RMS 31
2.3.6.1 RMSウィザード 32
2.3.6.2 プロセス監視機構 32
2.3.7 PAS 32
2.3.8 GDS 32
2.3.9 GFS 34
2.3.9.1 GFS共用ファイルシステム 34
2.3.9.2 メリット 36
2.3.10 GLS 36
2.3.10.1 高速切替方式 37
2.3.10.2 NIC切替方式 37
2.3.10.3 仮想NIC方式（Solaris） 38
2.3.10.4 仮想NIC方式（Linux） 38
2.3.10.5 GS/SURE連携方式（Solaris）、GS連携方式(Linux) 39

第3章 クラスタインタコネクトの詳細 40
3.1 概要 40
3.1.1 クラスタインタコネクトと通常のネットワークとの違い 40
3.1.2 インタコネクトプロトコル 40
3.2 クラスタインタコネクトの要件 40
3.2.1 冗長化 41
3.2.2 経路 41
3.2.2.1 ハートビート 42

3.2.3 設計時の検討項目 42
3.2.3.1 帯域幅 42
3.2.3.2 応答待ち時間 (レイテンシ) 43
3.2.3.3 信頼性 44
3.2.3.4 デバイスインタフェース (Solaris) 44
3.2.3.5 セキュリティ 44

第4章 RMS (Reliant Monitor Services) 45
4.1 RMSの概要 45
4.1.1 冗長化 45
4.1.2 アプリケーションの切替え 46
4.1.2.1 自動切替え 46
4.1.2.2 手動切替え 46
4.1.2.3 IPエイリアス 47
4.1.2.4 データの整合性 47
4.2 RMSの監視と切替え 47
4.2.1 BM (ベースモニタ) 47
4.2.2 構成定義ファイル 47
4.2.2.1 オブジェクト間の依存関係 47
4.2.2.2 オブジェクトタイプ (Object Type) 48
4.2.2.3 オブジェクト定義 (Object Definition) 48
4.2.3 構成スクリプト 49
4.2.4 ディテクタ 49
4.2.5 RMS環境変数 50
4.3 RMSの管理 50
4.4 カスタマイズオプション 50
4.4.1 汎用リソースタイプとディテクタ 50

第5章 RMSウィザード 51
5.1 RMSウィザードの概要 51
5.2 RMSウィザードのアーキテクチャ 51
5.3 RMS Wizard Tools 51
5.3.1 共用ディスク装置への対応 52

付録A リリース情報 53

用語集 54

索引 70

第1章 クラスタリングテクノロジーの概要

本章では、PRIMECLUSTER の主なコンポーネントを含むクラスタリングテクノロジーの基本概念とメリットについて説明します。

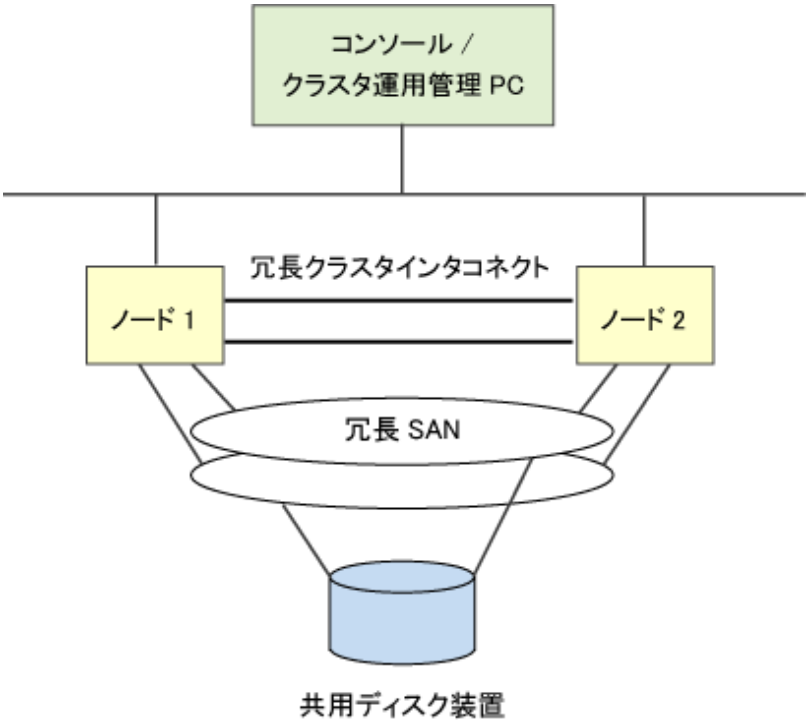
1.1 概要

一般にクラスタとは、以下の機能を提供するコンピュータ、またはコンピュータのパーティション (この単位をノードと呼びます) の結合体を意味します。

- 高可用性(HA)
  クラスタを構成する各コンポーネントの冗長化により実現する。
- 拡張性
  アプリケーションリソースを複数個動作させることにより実現する。

本書では、PRIMECLUSTERソフトウェア製品群により実現する、高可用性および拡張性を保証するクラスタシステムを中心に説明します。管理用クラスタシステムやR&D向け並列計算用クラスタなどの、他の種類のクラスタシステムについては本書の説明対象外です。以下の図は、一般的な2ノードクラスタシステムを示しています。

図1.1 一般的な2ノードクラスタシステム



PRIMECLUSTERでは、1ノードから構成されるシングルノードクラスタ構成もサポートしています。シングルノードクラスタでは、アプリケーションの状態を監視します。異常を検出した場合、アプリケーションを自動的に再起動し、復旧を試みることで、可用性を向上させることができます。

また、PRIMECLUSTER では、以下の仮想化環境にも対応しています。

- KVM 環境
- RHOSP 環境
- K5環境
- VMware 環境
- Oracle VM Server for SPARC 環境

- Oracle Solarisゾーン環境
  - カーネルゾーン
  - ノングローバルゾーン

1.2 高可用性 (HA)

HAクラスタは冗長化されたコンポーネントにより、各種の障害に対して対応することができます。PRIMECLUSTERを構成する各ノードは同じクラスタの他のノードとクラスタインタコネクトを利用して定周期間隔で通信を行いその応答を確認することにより、各ノードが稼動中かどうかを監視します。この定周期間隔の通信をハートビートと呼びます。

1.2.1 クラスタインタコネクト

クラスタインタコネクトとはPRIMECLUSTERがノード間の通信処理に使用する専用のネットワーク接続であり、クラスタシステムのもっとも基本的な構成要素です。クラスタインタコネクトの故障によるクラスタ全停止を防ぐために、クラスタインタコネクトの冗長化をぜひお勧めします。

クラスタインタコネクトは、ハートビート要求の他、各種のイベント通知、プロセス間通信、クラスタファイルアクセスなどのノード間のメッセージを伝送します。詳細は、本マニュアルの“第3章 クラスタインタコネクトの詳細”で説明します。

1.2.2 HAマネージャ

PRIMECLUSTER の HA マネージャは、クラスタ内のアプリケーションの高可用性を実現する Reliant Monitor Services (RMS) のことです。ユーザ業務が動作するための各種コンポーネントおよび、ユーザ業務が使用するリソースの状態を監視しています。ユーザ資産の整合性を保証し、ユーザ業務のリカバリを可能にするウィザードを提供します。

1.2.2.1 データ整合性の保証

以下の処理により、ユーザ資産であるデータ整合性を保証します。

- ユーザ業務の監視
- クラスタパーティションに対する処理
- 全てのクラスタノードの状態を確認した上でユーザ業務を自動起動する (RMS環境変数の設定により、制御されている場合は除く)

以下に、データ整合性を保証するための機能や処理の内容について説明します。

ユーザ業務の監視

RMSはアプリケーション固有のルールおよびクラスタ構成で設定されます。RMSの構成情報は、ユーザが定義するユーザ業務固有の定義と、動作するクラスタ環境の情報で構成されます。ディテクタが障害を検出すると、RMSは定義に従い適切な処置をとり、ユーザ業務を続行させるために必要なリソースのリカバリを行います。リカバリ処理はユーザ業務、およびリソースごとに定義することができます。

RMSには以下のリカバリ処理があります。

- ローカルリカバリ
  ユーザ業務を他のノードに切り替えずに、現在のノードで再度Onlineに戻すリカバリ処理
- リモートリカバリ
  ユーザ業務を他のノードに切り替える (フェイルオーバ)

クラスタパーティションに対する処理

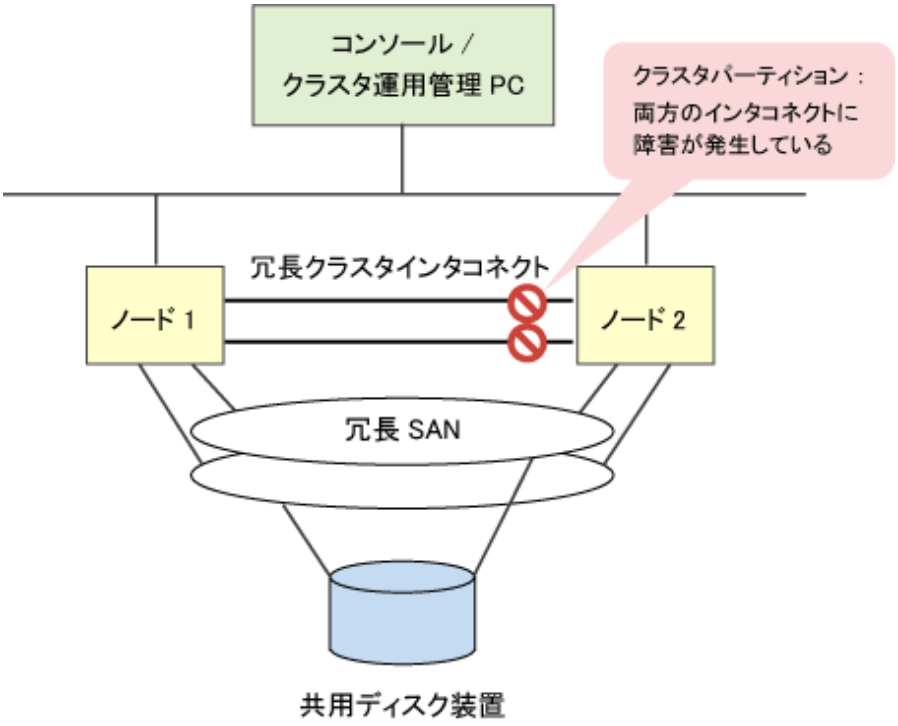
クラスタパーティションとはクラスタインタコネクトの障害により起こりうる現象のことです。

クラスタインタコネクトに障害が発生しても、クラスタの一部または全てのノードは処理を続行できますが、クラスタノードの一部のノード間通信は停止した状態になります。(スプリットブレイン状態とも呼ばれます。)スプリットブレイン状態を回避するには、クラスタインタコネクトの冗長化が有効ですが、充分ではありません。

以下の図は、冗長化したクラスタインタコネクトの両方の接続が切断したことにより、ノード1とノード2の通信が停止した場合の例です。

2つのノードはまだSANにアクセスすることができるため、各ノード上で独立してリカバリ処理を行うと、ユーザ業務がクラスタの2つのノードで互いに認識されないまま実行される可能性があります。この状態で互いに連携されていないユーザ業務が個別にデータを更新すると、共用ディスク装置上のユーザ資産を破損する危険性が生じます。

図1.2 2ノードクラスタのクラスタパーティション



ユーザ資産を破壊しないため、PRIMECLUSTERでは、以下のようなノード間の整合性を保つしくみを提供します。

1. クラスタシステム内の各ノードは、ハートビートによって相手ノードと通信できなかった場合（動作しているのか、または停止しているのかが不明な場合）、相手ノードをLEFTCLUSTER 状態に設定します。
2. LEFTCLUSTER 状態を解消します。
3. PRIMECLUSTERは、各ノードのリカバリ処理を開始する前に、クラスタシステム内のノードが以下の状態であることを確認します。
   - すべてノードが、動作中(UP)または停止中(DOWN)のいずれかの状態であること
     （LEFTCLUSTER状態のノードが存在しないこと）
   - 動作中のノードが他のすべての動作中のノードと通信可能であること

   PRIMECLUSTERでは、上記のようにノード間の整合性が保たれている状態を「クラスタ整合状態（クォーラム）」といいます。

PRIMECLUSTER のマニュアルでは「クラスタ整合状態」と「クォーラム」とは同じ意味です。クラスタ整合状態とは、クラスタの全てのノードが動作中 (UP) または 停止中(DOWN) のいずれかの状態で、動作中のUPノードが他の全てのUP状態のノードと通信可能な状態である場合に設定されます。クラスタ内で定義されているユーザ業務は、共用ディスク装置上のデータの変更を伴う処理を開始する前にクラスタがクラスタ整合状態になっていることを確認する必要があります。RMSはクラスタシステム内のユーザ業務起動前に、クラスタシステムがクラスタ整合状態になっていることを確認してから動作します。

PRIMECLUSTERは、クラスタシステムを構成するノードのアーキテクチャに応じた方法で、クラスタノードの強制停止を行います。PRIMECLUSTERはノードがLEFTCLUSTER状態であると判断すると、ノードを強制停止してユーザ業務のリカバリ処理 (ローカルリカバリまたはリモートリカバリ (フェイルオーバ)) を行い、データの整合性を保証します。

注意

クォーラムという用語の意味は、クラスタパーティションの処理を説明する文書にはさまざまな意味に用いられています。通常は、クラスタシステムを構成するノードが*n*個存在した場合、互いに (*n* + 1)/2 個のノードが参照できればクォーラムであり、クォーラムでないノードはI/O処理を行うことができません。PRIMECLUSTERではクォーラムの意味が上記の意味と異なるため、「クラスタ整合状態」という言葉を採用しています。

クラスタ整合性モニタ (CIM)

PRIMECLUSTERはクラスタ整合性モニタ (CIM) により、ユーザ業務がクラスタの複数ノードで共用されている資源を使った処理を、処理の競合をおこすことなく安全に処理することができるかどうかを判断します。つまり、処理を行うノードが、クラスタ整合状態であるクラスタシステムのメンバである場合、共用リソースを安全に使用することができることになります。

PRIMECLUSTERシステムにおける整合状態は、CIMが監視するクラスタシステムの全てのノードが動作中 (UP) または停止中 (DOWN) のいずれかの状態、かつ安全な状態である場合に設定されます。CIMが監視するノードは、CIM構成時に設定されたノード全てです。CIMはクラスタの状態を調べる場合、これらのノードのみを対象とします。

CIMは他のノードが安全である場合、クラスタ整合状態であると判断します。

クラスタを構成するノードの状態を調べる方式はCIM方式と呼ばれます。CIMは複数の異なるCIM方式を使用することができます。PRIMECLUSTERでは以下の方式が使用可能です。

- NSM
  ノード状態モニタ (NSM) はノードの状態を定周期で監視し、現在および過去のクラスタノードのノード状態を管理します。この方法はNULL方式またはデフォルトCIM方式とも呼ばれます。NSMはPRIMECLUSTER CFに組み込まれています。

- RCI
  RCI (Remote Cabinet Interface) は、Solarisシステム上でシステム間の状態通知やシステム制御を非同期で行う SPARC Enterprise M シリーズ専用制御機構です (詳細については、“PRIMECLUSTER Cluster Foundation 導入運用手引書” を参照してください)。

- XSCF SNMP
  XSCF SNMP (eXtended System Control Facility Simple Network Management Protocol) は、Solaris システム上でシステム間の状態通知やシステム制御を非同期で行う SPARC M10、M12 専用制御機構です (詳細については、“PRIMECLUSTER Cluster Foundation 導入運用手引書” を参照してください)。

- MMB
  MMB (Management Board) は、Linuxシステム上でシステム間の状態通知やシステム制御を非同期で行うPRIMEQUEST専用制御機構です (詳細については、“PRIMECLUSTER Cluster Foundation 導入運用手引書 ” を参照してください)。

PRIMECLUSTER は、複数の CIM 方式を登録して使用することができます。
複数のCIM方式が登録されている場合は、優先度の高い方式でノードの状態が判断できない場合にのみ優先度の低い方式を使用して確認します。例として、CIM方式としてRCIとNSMが登録され、RCI の方が優先度が高い場合は、CIMは、RCIを使用したCIM方式で確認を行います。
対象がノードまたはパーティションであれば、RCI CIM方式がUPまたはDOWNを返して処理は終了します。一方、RCI 方式によりチェックされるノードがRCIに接続されていない、またはRCIが故障していた場合は、RCI方式は失敗するため、CIMはNSMによるCIM方式を使用してノード状態を調べます。
PRIMEQUEST ノードでは、CIM 方式として MMB と NSM が使用され、MMB の方が優先度が高い場合は、CIM は、MMB を使用した CIM 方式で確認を行い、MMB CIM 方式が UP または DOWN を返して処理は終了します。MMB 方式によりチェックされるノードが MMB に接続されていない、または MMB が故障していた場合は、MMB 方式は失敗するため、CIM は NSM による CIM 方式を使用してノード状態を調べます。

CIMは対象ノードに関して、クラスタ整合状態である (TRUE)、またはクラスタ整合状態でない (FALSE) のいずれかのノード状態を通知します。TRUEとFALSEの定義は以下のとおりです。

- TRUE
  全てのCIMノードにとって、UP、またはDOWNの状態が既知の状態

- FALSE
  全てのCIMノードにとって、UP、またはDOWNの状態が不明な状態

シャットダウン機構 (PRIMECLUSTER SF)

CIMは、クラスタ整合状態である場合にユーザ業務に対して動作することを許可しますが、クラスタ整合状態でない場合はこれを解決するような処理を行いません。高可用性要件ではクラスタ整合状態を保証するために複数の方式が使用されます。しかし、ノード間の協調を必要とせず、かつ完全に効果のある方法は1つだけです。PRIMECLUSTERでは、クラスタ整合状態を妨げるような問題が発生した場合は、シャットダウン機構 (SF) を使用して、クラスタ整合状態に戻します。図1.2 2ノードクラスタのクラスタパーティションの例では、2つのノードは互いに相手ノードに対してLEFTCLUSTERを通知した結果、CIMはFALSEと判断します。PRIMECLUSTERは、クラスタシステムをクラスタ整合状態にするため、SFは強制的に相手ノードを停止することで、生存ノードを1つにして競合の発生しない安全な状態にします。

SFの設定により、PRIMECLUSTERは異常となったノードを強制停止することができます。SFはノードを強制停止するような要求を受けると、ノードの強制停止を行い、成功した場合にノードの状態はLEFTCLUSTERからDOWNに変化します。

状態をLEFTCLUSTERからDOWNに変更するとPRIMECLUSTERは各種、リカバリ処理を開始します。ノードの強制停止の方法は、システムによって異なります。たとえば、Solarisでは有効なシャットダウンエージェントが、Linuxでは使用できないことがあります。

システムに登録されている全ての方式を実行しても要求したノードの強制停止成功の応答が得られない場合、処理はそこで停止します。この場合、クラスタはクラスタ整合状態ではない状態のままなので、オペレータによる操作が必要になります。

この方式により、誤ってクラスタパーティションに分割されたクラスタシステムの2箇所でユーザ業務を実行し、その競合によりユーザデータが破壊されることを防ぐことができます。また、システム負荷 (System Load) が著しく高いことなどが原因で、ノードがハートビートに対する応答に失敗し、後から復活するという状況でも、ユーザ資産は競合から保護されます。

注意

PRIMECLUSTERはハードウェア固有のさまざまな方法で、SolarisまたはLinuxが稼動するノードをリセットするように定義することができます。詳細については、“PRIMECLUSTER Cluster Foundation 導入運用手引書” を参照してください。

非同期監視 (Monitoring Agent)

PRIMECLUSTERは、ハードウェアの機能を使用してシステム状態の変化をすばやく検出し、クラスタを構成するノードに通知します。PRIMECLUSTERのこの監視機能を非同期監視(Monitoring Agent)といいます。非同期監視を使用しない場合、ノードのパニックを検出する機能はクラスタのハートビートタイムアウトのみであるため、デフォルトのハートビート間隔の設定では検出に10秒が必要です。非同期監視を使用した場合、ノードのパニックを即時に検出することができます。非同期監視テクノロジーにより、PRIMECLUSTERは監視対象ノードの障害からすばやく復旧することができます。非同期監視は、シャットダウン機構のプラグインとして実装されています。

ノードに異常が発生した場合、PRIMECLUSTERは以下の処理を行います。

1. ノード異常の検出
2. 異常となった障害の通知
3. ノード状態の確認
4. ノードの強制停止

MAはノード異常を検出すると、ただちにSFに通知します。SFは、MAによる障害通知が本当に正しいかどうかを判断するために、ノード状態に関する冗長確認を行います。この検証処理は、正常に動作しているノードが誤って停止されないようにするために行われます。

SFは、以下のようにしてノードの状態を確認します。

- 全ての登録済みMAから再度ノードの状態情報を採取する。
- CFハートビート要求への応答があったかどうかを確認する。

全てのMAからSFに対してノード障害の通知があり、CFからSFに対して、ハートビート要求への応答がなかった旨の通知があった場合は、SFはMAに対して障害の発生したノードの強制停止を要求します。ノードの強制停止が完了すると、他方のノードはDOWNの状態になります。

I/Oフェンシング機能

共有ディスク装置が接続されたクラスタ構成では、SCSI-3 Persistent Reservation による排他制御機能を利用し、両ノードからの同時アクセスを防止します。

本機能は、以下の仮想化環境でのみ使用できます。

- VMware 環境
- Oracle VM Server for SPARC 環境

1.2.2.2 ウィザード

ユーザ業務を適切にリカバリするには、ユーザ業務の正常な動作に必要なリソースをあらかじめRMSに定義し、その状態を通知しておかなければなりません。リソース構成およびリソース間の関係はきわめて複雑になる場合があります。RMS Wizard Toolsは、これらの情報をRMSに指定するための構成定義を行います。また、userApplication Configuration Wizardを用いることでこれらをGUIにより構成することができます。RMS Wizard Toolsは、クラスタや一般的なアプリケーションサービスに関係する一般的な情報を設定します。

1.2.3 パトロール診断機能 (Solaris)

パトロール診断機能とは、待機ノードに接続された以下のハードウェアを定期的に診断する機能です。

- 共用ディスク装置

  電源オフ、ケーブル抜け(アダプタ側・装置側)などにより、共用ディスク装置が使用できなくなっていないかを診断します。

  診断の結果、共用ディスク装置に異常を検出した場合、エラーメッセージを出力します。

- ネットワークインタフェースカード

  ケーブル抜けなどにより、ネットワークインタフェースカードが通信できなくなっていないかを診断します。

  診断の結果、ネットワークインタフェースカードの異常を検出した場合、エラーメッセージを出力し、待機ノードへの切り替え(フェイルオーバ)をできないようにします。

1.3 拡張性 (スケーラビリティ)

高い拡張性もPRIMECLUSTERの特長の1つです。PRIMECLUSTERの拡張性は、クラスタの処理能力の拡張によって実現されます。拡張性が重要なユーザ業務の形態は、基本的に以下の2種類のタイプがあります。

- クラスタソフトウェアと密接に連携した分散環境向け
- クラスタを意識しないもの

クラスタソフトウェアと連携するケース

クラスタソフトウェアと連携する拡張性のあるアプリケーションの一例として、Oracle RACがあります。Oracle RACはクラスタシステム上の一部または全てのノード上で、データベースサーバのOracleインスタンスを起動します。

クラスタを意識しないもの

クラスタ環境を意識せずに動作することができるようなユーザ業務は、複数ノードで同時に動作させることができます。同じファイルにアクセスするアプリケーションを複数同時に実行するような場合、Global File Services (以降、GFS) の機能によりクラスタノード間でファイルを共用し、複数ノードでアプリケーションを動作させることで負荷分散効果を高めることができます。

参照

GFSの詳細については本マニュアルの “2.3 PRIMECLUSTERのモジュール” および “PRIMECLUSTER Global File Services 説明書” を参照してください。

1.4 シングルノードクラスタ

シングルノードクラスタとは、1 ノードで構成されるクラスタシステムで、そのノード上の業務を監視・制御することができます。

異常を検出した場合、アプリケーションを自動的に再起動し、復旧を試みることで、可用性を向上させることができます。

また、クラスタアプリケーションの作成、テストを行うための開発環境としても利用できます。

ただし、ハードウェア故障が発生した場合、業務停止となります。また、シングルノードクラスタ運用ではフェイルオーバすることはありません。

1.5 ノード間のデータ引継ぎ

PRIMECLUSTERでは、クラスタのノード間でデータを引き継ぐ方法として、以下の方式が選択できます。

- 共用ディスク方式

  共用ディスク装置にデータを格納する方式です。

  以下の共用ディスク装置が使用可能です。

- ー SAN (Storage Area Network)を経由して複数のサーバに共用接続されたディスク装置

  データアクセス性能やデータの可用性が重視される場合、またはデータ容量の拡張性が必要な場合に適しています。

- ー NAS (Network Attached Storage)装置

  比較的に低コストで構築できます。

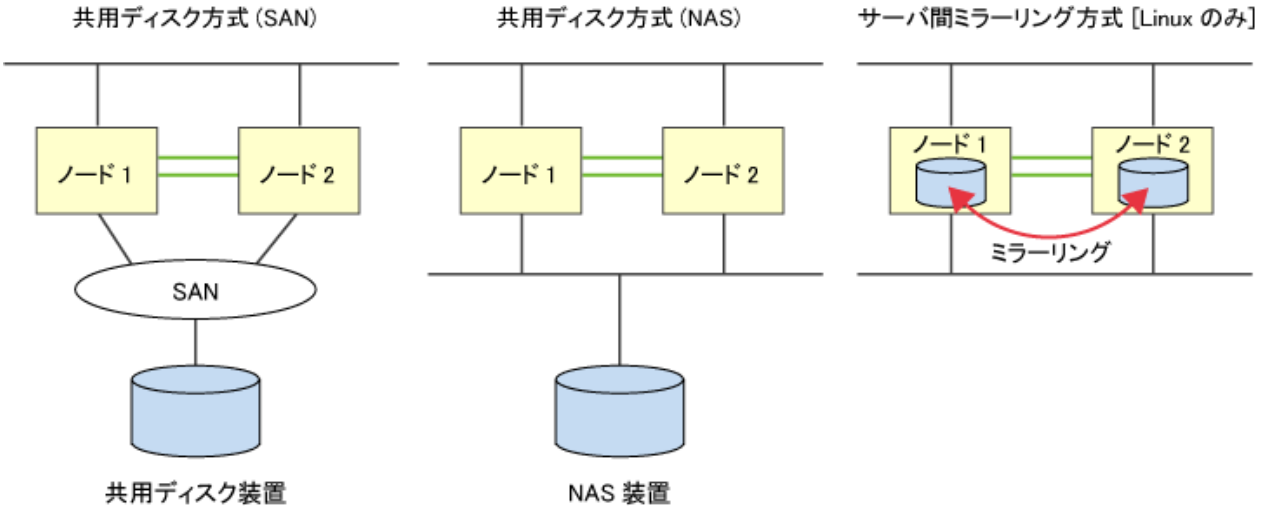
- サーバ間ミラーリング方式(Linux)

  各ノードのローカルディスクにデータを格納し、Global Disk Services（以降、GDS）のサーバ間ミラーリング機能によって、それらのローカルディスクどうしをネットワーク経由でミラーリングする方式です。

  高価な外部ストレージを必要としないため、低コストで構築できます。

  データ量とデータ更新量が少ない小規模システムに適しています。

図1.3 ノード間のデータ引継ぎ



1.6 仮想化対応

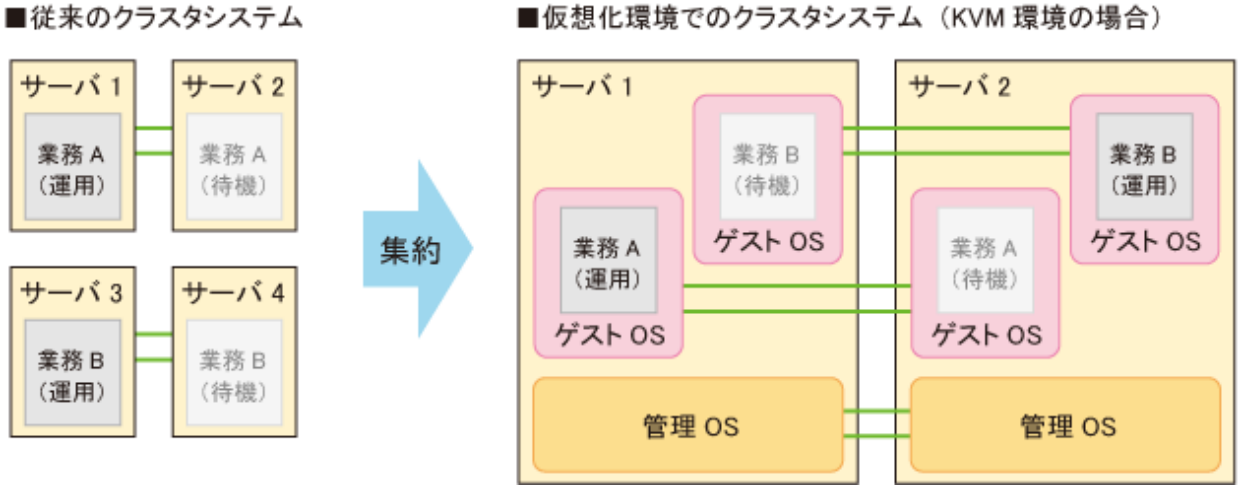
PRIMECLUSTERは、以下の仮想化環境で冗長構成を可能にして、集約されたシステムの高信頼化を実現します。

- KVM環境
- RHOSP 環境
- K5環境
- VMware環境
- Oracle VM Server for SPARC環境
- Oracle Solaris ゾーン環境
  - ー カーネルゾーン
  - ー ノングローバルゾーン

従来は、業務単位に異なるサーバで、クラスタシステムを構築/運用していましたが、仮想化環境に対応したことにより、複数の業務を1つのサーバに集約することが可能になります。

サーバ内では、業務単位にCPUリソースを配分して運用することができます。

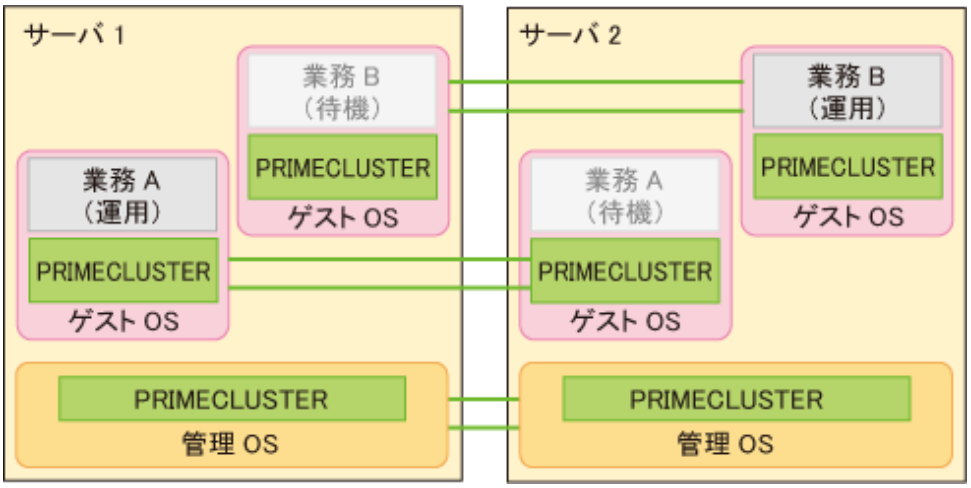
図1.4 従来のクラスタシステムと仮想化環境でのクラスタシステム（KVM環境の場合）



KVM環境 の場合

管理OS/ゲストOSそれぞれにPRIMECLUSTERを導入することで、アプリケーションの異常時だけでなく、OS異常時にも迅速な切替えが可能です。

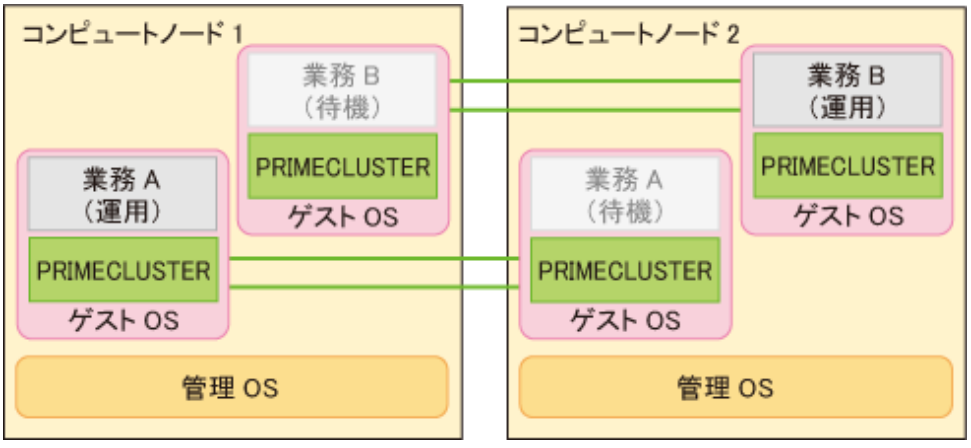
図1.5 仮想化環境でのクラスタシステム（KVM環境の場合）



RHOSP環境 の場合

ゲストOSにPRIMECLUSTERを導入することで、アプリケーションやOS異常時に迅速な切替えが可能です。

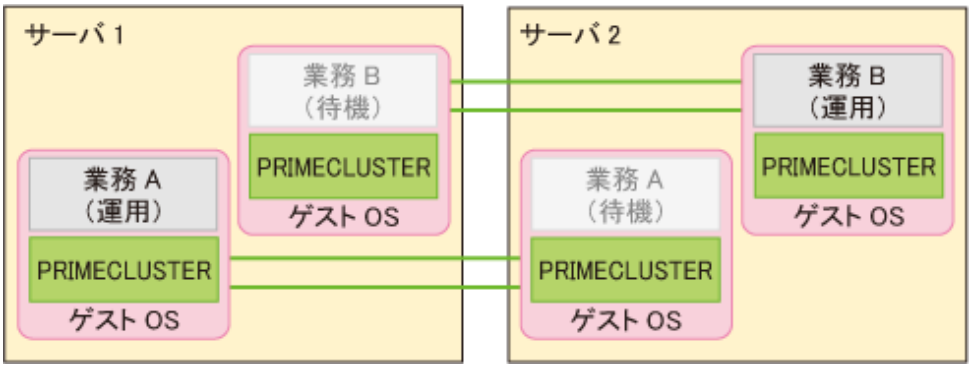
図1.6 仮想化環境でのクラスタシステム（RHOSP環境の場合）



K5環境 の場合

ゲストOSにPRIMECLUSTERを導入することで、アプリケーションやOS異常時に迅速な切替えが可能です。

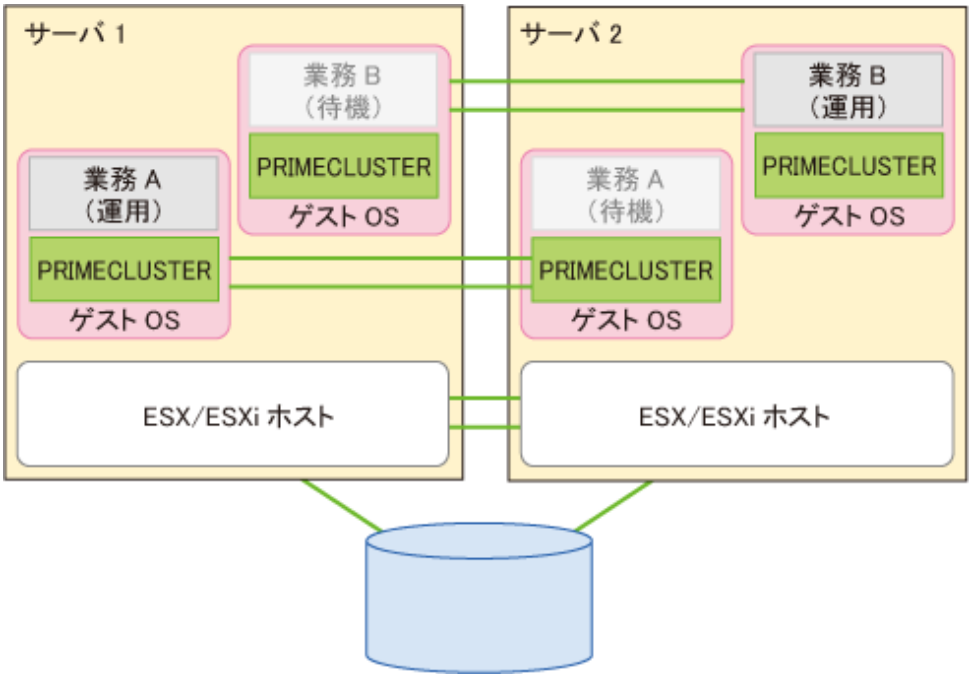
図1.7 仮想化環境でのクラスタシステム（K5環境の場合）



VMware環境の場合

ゲストOSにのみ、PRIMECLUSTERを導入します。
アプリケーションやOSの異常時には、VMware vCenter Server連携機能、または、共有ディスクを使用したI/Oフェンシング機能により、安全で確実な切替えが可能です。

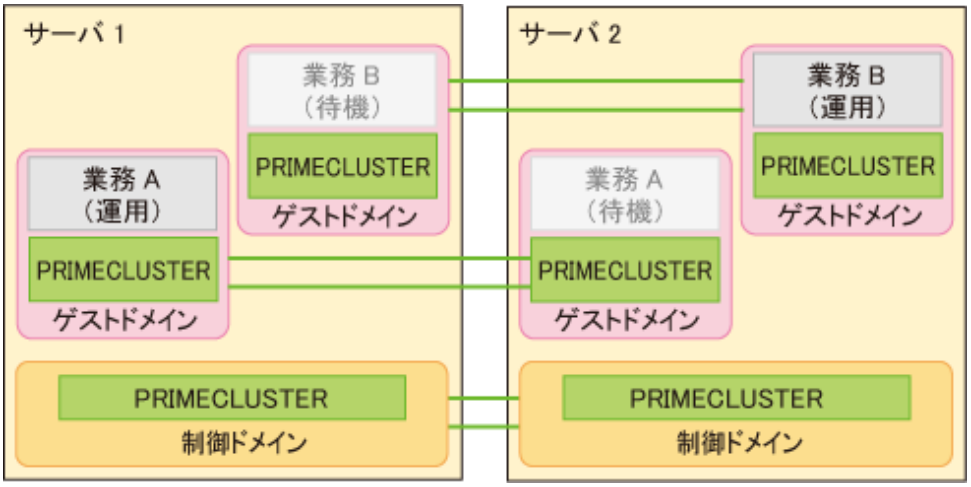
図1.8 仮想化環境でのクラスタシステム（VMware環境の場合）



Oracle VM Server for SPARC環境の場合

制御ドメイン/ゲストドメインそれぞれにPRIMECLUSTERを導入することで、ゲストドメインの異常時だけでなく、パーティションの 異常時にも迅速な切替えが可能です。

図1.9 仮想化環境でのクラスタシステム（Oracle VM Server for SPARC環境の場合）



Oracle Solarisカーネルゾーン環境の場合

制御ドメインまたは、ゲストドメインにカーネルゾーンを作成し、制御ドメインまたは、ゲストドメインにもPRIMECLUSTER を導入することで、ドメインの異常時だけでなく、パーティション異常時にも迅速な切替えが可能です。

図1.10 仮想化環境でのクラスタシステム（制御ドメイン上のOracle Solarisカーネルゾーン環境の場合）

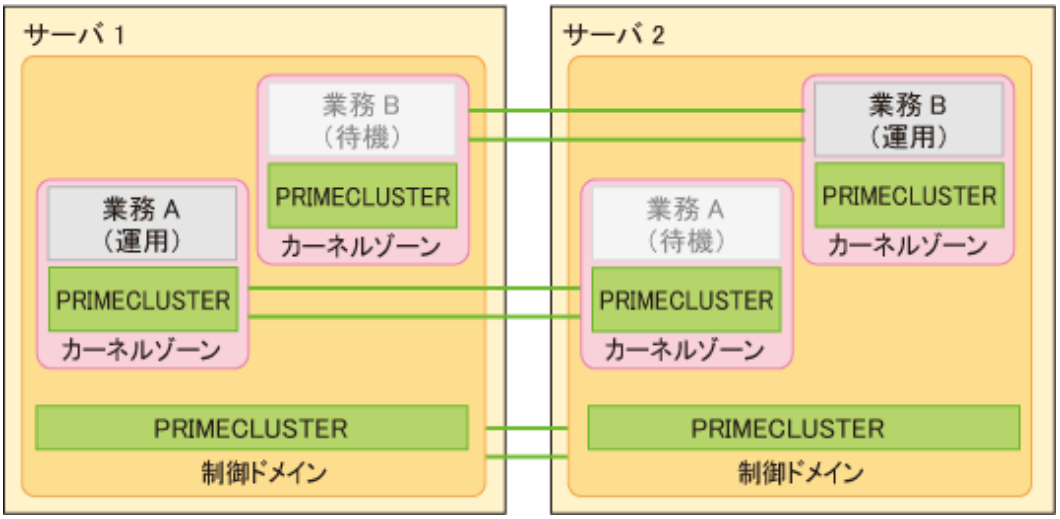
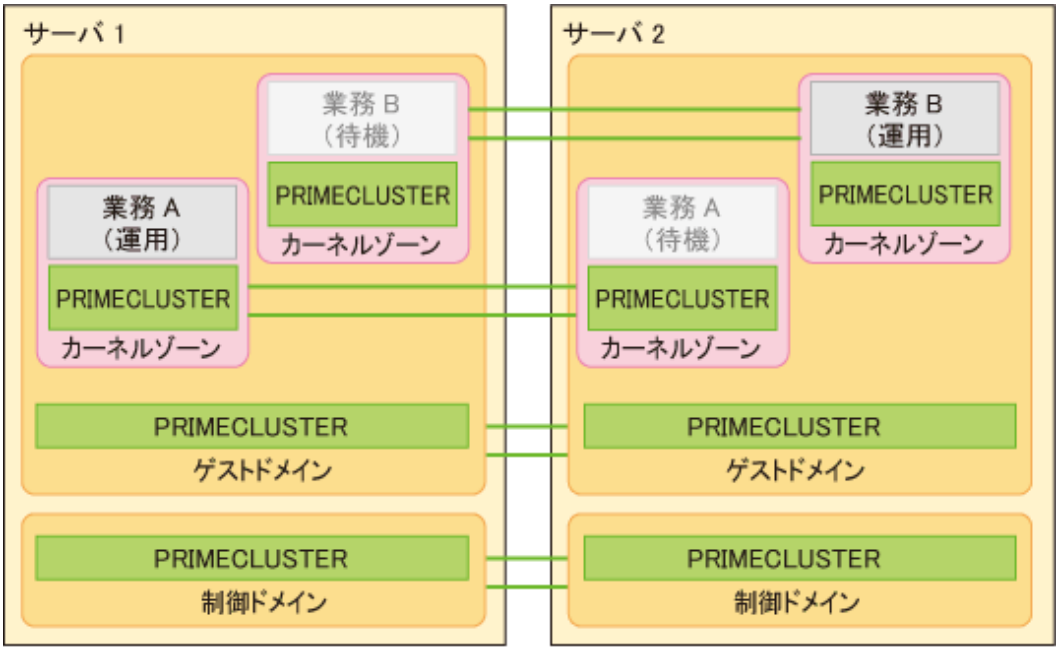


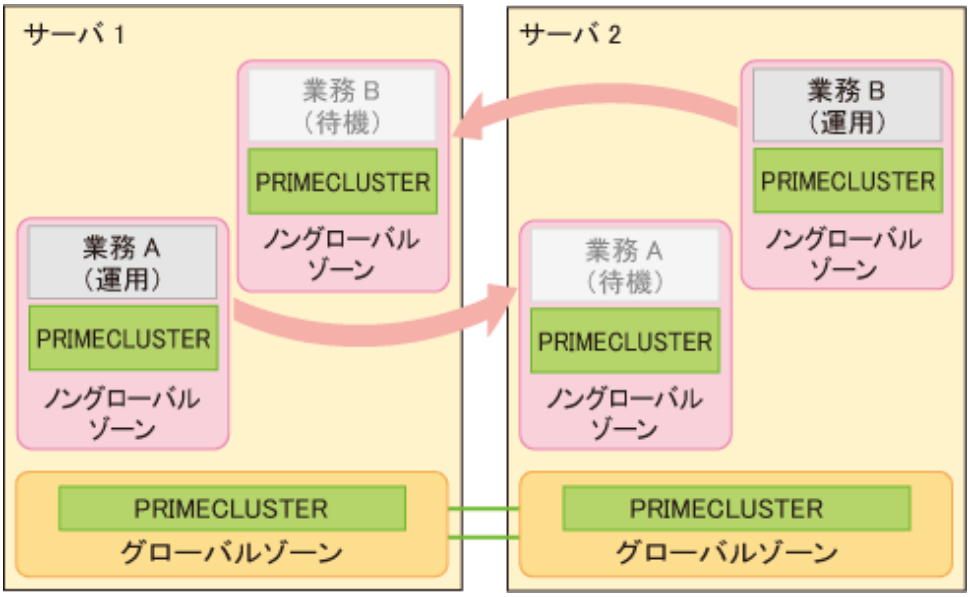
図1.11 仮想化環境でのクラスタシステム（ゲストドメイン上のOracle Solarisカーネルゾーン環境の場合）



Oracle Solarisノングローバルゾーン環境の場合

グローバルゾーンにPRIMECLUSTERを導入し、グローバルゾーン異常時にサーバ単位での切替えが可能です。
また、ノングローバルゾーンにもPRIMECLUSTERを導入することで、アプリケーション異常時にもノングローバルゾーン毎の切替えが可能です。

図1.12 仮想化環境でのクラスタシステム（Oracle Solarisノングローバルゾーン環境の場合）



1.7 異常発生時の切替え動作

ここでは、クラスタシステム構成別の異常発生時の切替え動作について説明します。

1.7.1 Linux

Linuxで、以下の環境の場合のクラスタシステムの可用性について説明します。

- 物理環境のクラスタシステム

- 管理OS異常切替機能を使用したクラスタシステム(KVM)
- 異なる管理OS上のゲストOS間クラスタ(KVM)
- 同一管理OS上のゲストOS間クラスタ(KVM)
- 異なるコンピュートノード上のゲストOS間クラスタ(RHOSP)
- 同一コンピュートノード上のゲストOS間クラスタ(RHOSP)
- ゲストOS間クラスタ(K5)
- 異なるESXiホスト上のゲストOS間クラスタ(VMware)
- 同一ESXiホスト上のゲストOS間クラスタ(VMware)

以下の表では、各監視対象の異常検出の可否についてまとめています。

表1.1 クラスタシステム構成別の可用性

監視対象	物理サーバ	KVM			RHOSP		K5	VMware	
		管理OS異常切替機能を使用したクラスタ	異なる管理OS上のゲストOS間クラスタ	同一管理OS上のゲストOS間クラスタ	異なるコンピュートノード上のゲストOS間クラスタ	同一コンピュートノード上のゲストOS間クラスタ	ゲストOS間クラスタ	異なるESXiホスト上のゲストOS間クラスタ	同一ESXiホスト上のゲストOS間クラスタ
1. 筐体	○	○	×	×	○*1	×	×	○*2	×
2. 共用ディスクおよびディスクアクセスパス	○	○	○	×	○	×	○	○	×
3. 業務LAN	○	○	○	×	○	×	○	○	×
4. OS（物理、管理OS/ESXiホスト）	○	○	×	×	○*1	×	×	○*2	×
5. OS（ゲストOS）	－	○	○	○	○	○	○	○*3	○*4
6. 業務（クラスタアプリケーション）	○	○	○	○	○	○	○	○	○

異常時の業務継続 ○：可、×：不可、－：対象外

*1 コンピュートインスタンスの高可用設定により業務継続可能
コンピュートインスタンスの高可用設定の詳細については、“Red Hat OpenStack Platform コンピュートインスタンスの高可用性”を参照してください

*2 I/Oフェンシング機能使用時、または、VMware vCenter Server連携機能とVMware vSphere HA使用時
ゲストOSのハングアップを検出しゲストOSを待機系に自動切替えできない場合は、LEFTCLUSTERとなります

*3 ゲストOSを待機系に自動切替えできない場合は、LEFTCLUSTERとなります

*4 VMware vCenter Server連携機能使用時のみ自動切替え可能となります

図1.13 物理環境

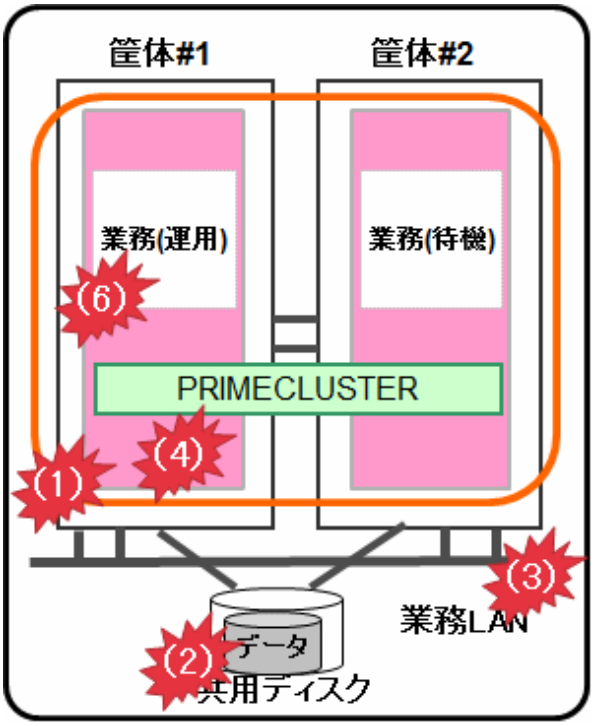
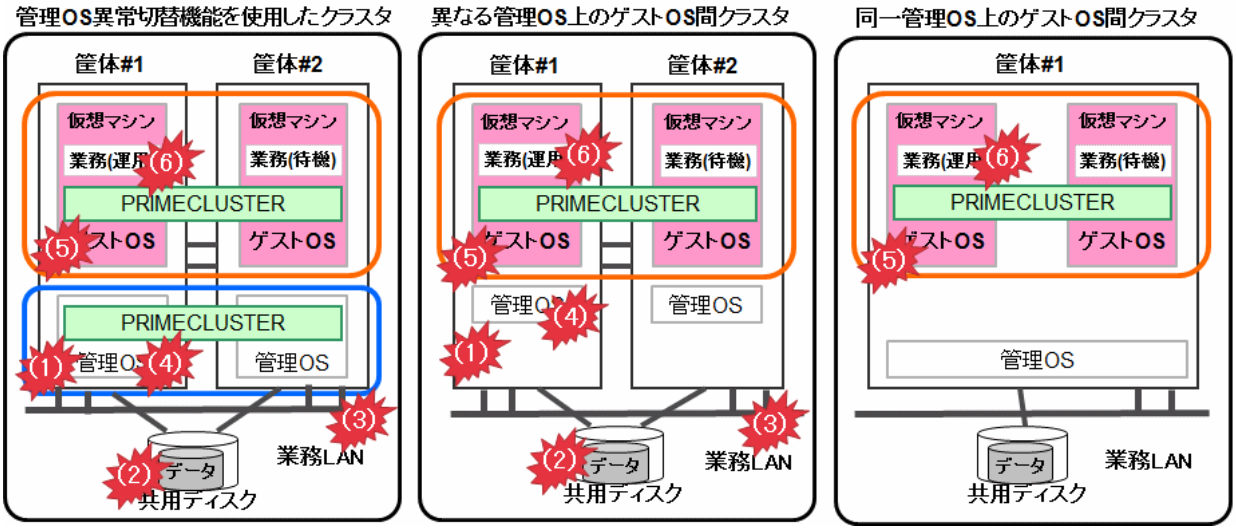


図1.14 仮想環境



RHOSP環境の場合、管理OSをコンピュートノードと、VMware環境の場合、管理OSをESXiホストと読み替えてください。K5環境の場合、異なる管理OS上のゲストOS間クラスタの場合の図と同様です。

監視対象の異常検出方法

1. 筐体

   PRIMEQUEST 2000の場合はサーバ管理ボード(MMB)、PRIMEQUEST 3000の場合はiRMC/MMBと連携した非同期監視機能が、CPUやメモリ等の異常を契機とするパニック、およびリセットを即時検出し、待機系に切り替えます。PRIMERGYおよび仮想環境の場合、ハートビート監視で異常を検出し、待機系に切り替えます。*1

2. 共用ディスクおよびディスクアクセスパス

   ボリューム管理機能(GDS)と組み合わせることで、ディスクアクセスおよび、ディスクアクセスパスの故障を検出(Gdsリソースで監視)し、ディスクアクセス不可または、ディスクアクセスパスの全系故障の場合に待機系に切り替えます。

3. 業務LAN

   ネットワーク多重化機能(Global Link Services。以降、GLS)と組み合わせることで、業務LANのネットワークアダプタや経路の故障を検出（Glsリソースで監視）し、ネットワークの全系故障の場合に待機系に切り替えます。

4. OS（物理、管理OS/ESXiホスト）

   ハートビート監視で異常を検出し、待機系に切り替えます。*1

5. OS（ゲストOS）

   ハートビート監視で異常を検出し、待機系に切り替えます。

6. 業務（クラスタアプリケーション）

   クラスタアプリケーションのリソース異常発生時に待機系に切り替えます。

*1 異なる管理OS上のゲスト間クラスタ（RHOSP、VMware）の場合、LEFTCLUSTERとなります。コンピュートインスタンスの高可用設定(RHOSP)やvSphere HA機能(VMware)により、ゲストOSが再起動することで、LEFTCLUSTER状態が自動的に解消され、待機系に切り替わります。

1.7.2 Oracle Solaris

Oracle Solarisで、以下の環境の場合のクラスタシステムの可用性について説明します。

- 物理環境のクラスタシステム
- Oracle VM Server for SPARC環境のクラスタシステム
- Oracle Solaris ゾーン環境のクラスタシステム
  - Oracle Solaris カーネルゾーン環境のクラスタシステム
  - Oracle Solaris ノングローバルゾーン環境のクラスタシステム

1.7.2.1 Oracle Solaris(物理、Oracle VM Server for SPARC環境)

Oracle Solarisで、以下の物理環境、Oracle VM Server for SPARC環境の場合のクラスタシステムの可用性について説明します。

- 物理環境のクラスタシステム
- Oracle VM Server for SPARC環境の異なる物理パーティション間のゲストドメインクラスタ
- Oracle VM Server for SPARC環境の同一物理パーティション内のゲストドメインクラスタ
- Oracle VM Server for SPARC環境の制御ドメイン間クラスタ

以下の表では、各監視対象の異常検出の可否についてまとめています。

表1.2 クラスタシステム構成別の可用性

監視対象	物理環境	Oracle VM Server for SPARC環境		
		異なる物理パーティション間のゲストドメインクラスタ	同一物理パーティション内のゲストドメインクラスタ	制御ドメイン間クラスタ
1. 物理パーティション	○	○	×	○
2. 共用ディスクおよびディスクアクセスパス	○	○	×	○
3. 業務LAN	○	○	×	○
4. OS(物理、制御ドメイン)	○	○	○*1	○
5. OS(ゲストドメイン)	－	○	○	○*2
6. 業務（クラスタアプリケーション）	○	○	○	○*3

異常時の業務継続 ○：可、×：不可、－：対象外

*1 制御ドメインのOS異常時もゲストドメインのOSは継続動作可のため業務継続可

*2 ゲストドメインのOSの監視不可。ゲストドメインの状態(ldm list-domainで表示される状態)の異常時、制御ドメインのPRIMECLUSTERで監視して待機系に切り替えて業務継続可

*3 制御ドメイン上の業務（クラスタアプリケーション）監視可、ゲストドメイン上の業務監視不可

図1.15 物理環境

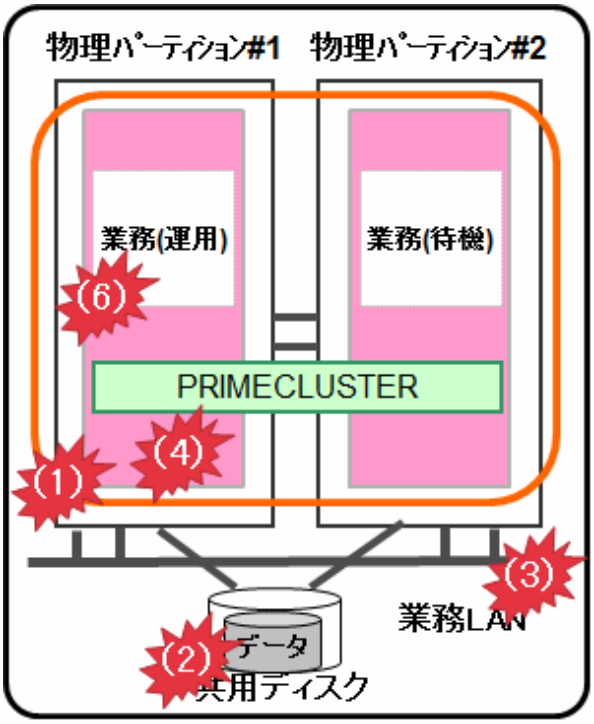
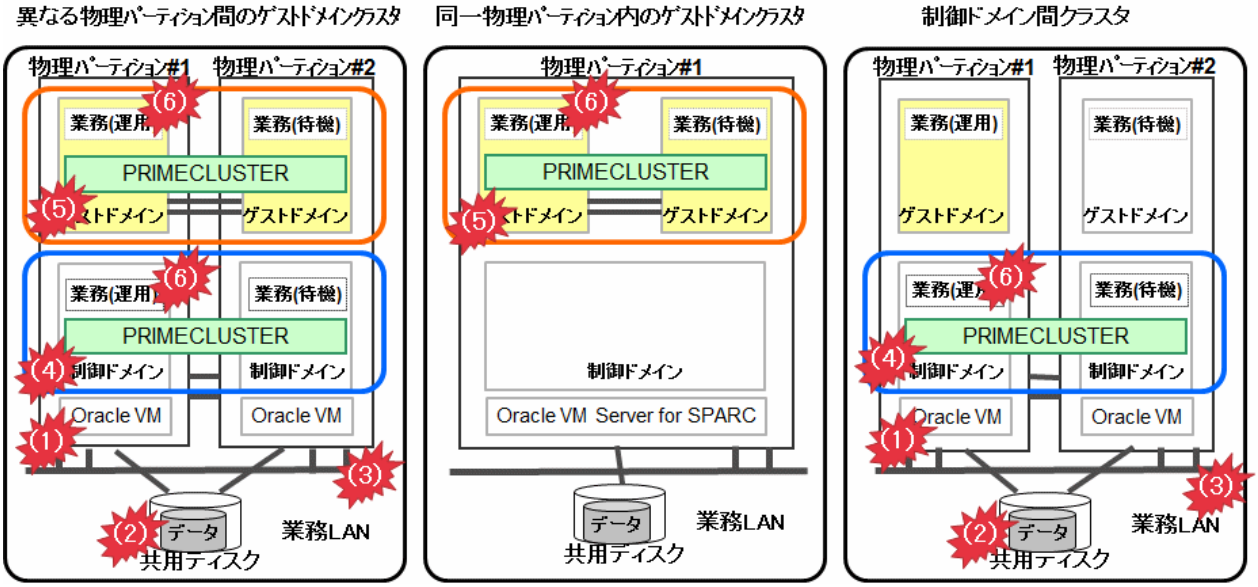


図1.16 Oracle VM Server for SPARC環境



監視対象の異常検出方法

1. 物理パーティション

   サーバのシステム監視機構と連携した非同期監視が、CPUやメモリ等の異常を契機とするパニック、およびリセットを即時検出し、待機系に切り替えます。

2. 共用ディスクおよびディスクアクセスパス

   ボリューム管理機能（GDS）と組み合わせることで、ディスクアクセスおよび、ディスクアクセスパスの故障を検出（Gdsリソースで監視）し、ディスクアクセス不可または、ディスクアクセスパスの全系故障の場合に待機系に切り替えます。

3. 業務LAN

   ネットワーク多重化機能(GLS)と組み合わせることで、業務LANのネットワークアダプタや経路の故障を検出（Glsリソースで監視）し、ネットワークの全系故障の場合に待機系に切り替えます。

4. OS（物理、制御ドメイン）

   非同期監視により、OSのパニック、およびリセットを即時検出し、待機系に切り替えます。クラスタインタコネクト(LAN)定周期監視によりOSのハングアップを検出し、待機系に切り替えます。

   同一物理パーティション内のゲストドメインクラスタの場合、制御ドメインのOS異常は検出できません。（制御ドメインがシングルのため）

5. OS(ゲストドメイン)

   非同期監視により、OSのパニック、およびリセットを即時検出し、待機系に切り替えます。クラスタインタコネクト(LAN)定周期監視によりOSのハングアップを検出し、待機系に切り替えます。

   制御ドメイン間クラスタの場合、ゲストドメインの業務の異常は検出できません。

6. 業務（クラスタアプリケーション）

   クラスタアプリケーションのリソース異常発生時に待機系に切り替えます。

1.7.2.2 Oracle Solaris(Oracle Solaris カーネルゾーン環境)

Oracle Solarisのカーネルゾーンで、以下の場合のクラスタシステムの可用性について説明します。

- 異なる物理パーティション間のカーネルゾーン間クラスタ(制御ドメイン)
- 異なる物理パーティション間のカーネルゾーン間クラスタ(ゲストドメイン)
- 同一物理パーティション内のカーネルゾーン間クラスタ(制御ドメイン)
- 同一物理パーティション内のカーネルゾーン間クラスタ(ゲストドメイン)

表1.3 クラスタシステム構成別の可用性

監視対象	Oracle Solaris カーネルゾーン環境			
	異なる物理パーティション間のカーネルゾーン間クラスタ (制御ドメイン)	異なる物理パーティション間のカーネルゾーン間クラスタ (ゲストドメイン)	同一物理パーティション内のカーネルゾーン間クラスタ (制御ドメイン)	同一物理パーティション内のカーネルゾーン間クラスタ (ゲストドメイン)
1. 物理パーティション	○	○	×	×
2. 共用ディスクおよびディスクアクセスパス	○	○	×	×
3. 業務LAN	○	○	×	×
4. OS(物理、制御ドメイン)	○	○*1	×	○*1
5. OS(ゲストドメイン)	－	○	－	○*2
6. OS(カーネルゾーン)	○	○	○	○
7. 業務（クラスタアプリケーション）	○	○	○	○

異常時の業務継続 ○：可、×：不可

*1 制御ドメインのOS異常時もゲストドメインのOSは継続動作可のため業務継続可

*2 同一ゲストドメイン内のカーネルゾーン間クラスタの場合は、業務継続不可

図1.17 Oracle Solaris カーネルゾーン環境(異なる物理パーティション間)

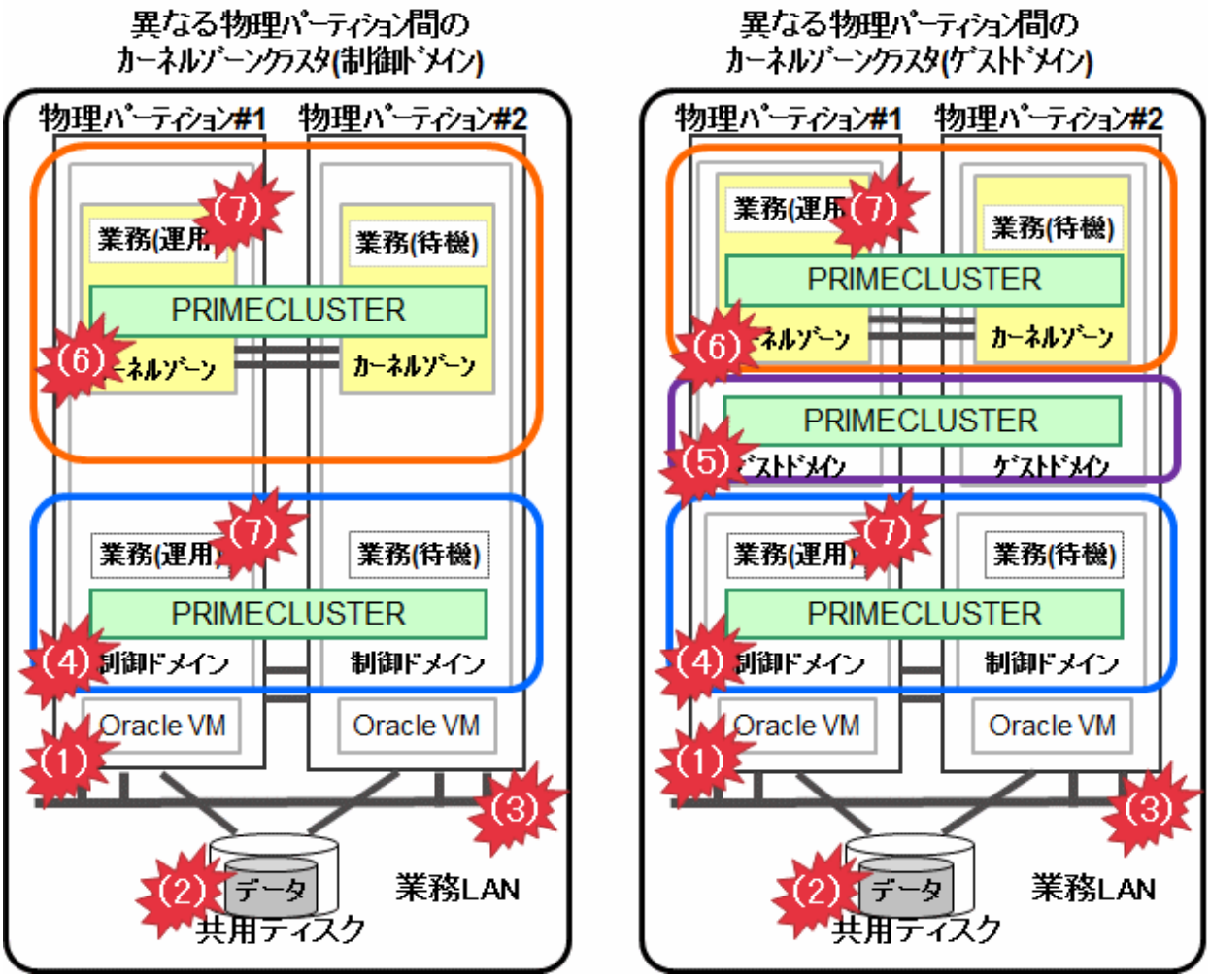
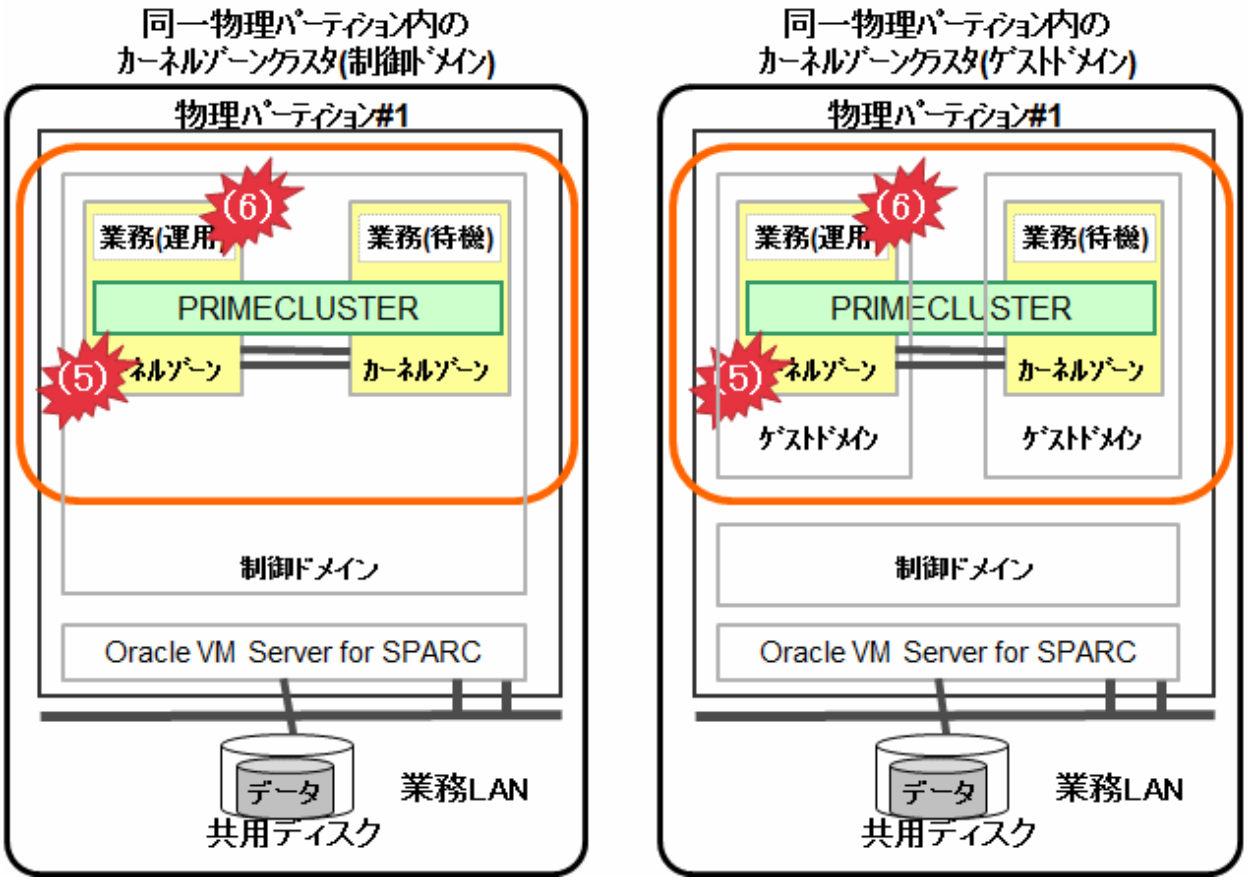


図1.18 Oracle Solaris カーネルゾーン環境(同一物理パーティション内)



監視対象の異常検出方法

1. 物理パーティション

   サーバのシステム監視機構と連携した非同期監視が、CPUやメモリ等の異常を契機とするパニック、およびリセットを即時検出し、待機系に切り替えます。

2. 共用ディスクおよびディスクアクセスパス

   ボリューム管理機能(GDS)と組み合わせることで、ディスクアクセスおよび、ディスクアクセスパスの故障を検出(Gdsリソースで監視)し、ディスクアクセス不可または、ディスクアクセスパスの全系故障の場合に待機系に切り替えます。

3. 業務LAN

   ネットワーク多重化機能(GLS)と組み合わせることで、業務LANのネットワークアダプタや経路の故障を検出(Glsリソースで監視)し、ネットワークの全系故障の場合に待機系に切り替えます。

4. OS(物理、制御ドメイン)

   非同期監視により、OSのパニック、およびリセットを即時検出し、待機系に切り替えます。また、クラスタインタコネクト(LAN)定周期監視によりOSのハングアップを検出し、待機系に切り替えます。

   同一物理パーティション内のカーネルゾーンクラスタの場合、制御ドメインのOS異常は検出できません。(制御ドメインがシングルのため)

5. OS(ゲストドメイン)

   非同期監視により、OSのパニック、およびリセットを即時検出し、待機系に切り替えます。また、クラスタインタコネクト(LAN)定周期監視によりOSのハングアップを検出し、待機系に切り替えます。

   同一ゲストドメイン内のカーネルゾーンクラスタの場合、ゲストドメインのOS異常は検出できません。(ゲストドメインがシングルのため)

6. OS(カーネルゾーン)

   クラスタインタコネクト(LAN)定周期監視によりOSのパニック、リセット、およびハングアップを検出し、待機系に切り替えます。

7. 業務(クラスタアプリケーション)

クラスタアプリケーションのリソース異常発生時に待機系に切り替えます。

1.7.2.3 Oracle Solaris(Oracle Solaris ノングローバルゾーン環境)

Oracle Solaris ノングローバルゾーンで、以下の環境の場合のクラスタシステムの可用性について説明します。

- コールドスタンバイ環境(待機側のノングローバルゾーンが起動していない状態(待機側の業務も起動していない状態))
- ウォームスタンバイ環境(待機側のノングローバルゾーンは起動している状態(待機側の業務は起動していない状態))
- シングルノードクラスタ環境

表1.4 クラスタシステム構成別の可用性

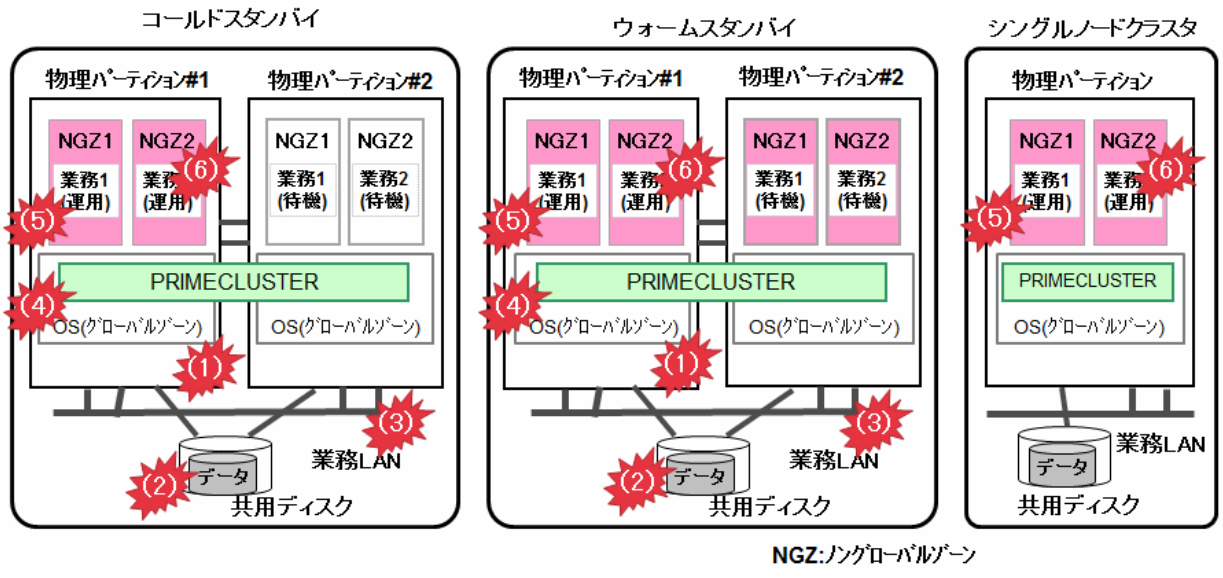
監視対象	コールドスタンバイ	ウォームスタンバイ	シングルノードクラスタ
1. 物理パーティション	○	○	―
2. 共用ディスクおよびディスクアクセスパス	○	○	―
3. 業務LAN	○	○	―
4. OS(グローバルゾーン)	○	○	―
5. OS(グローバルゾーン)	○	○	○*1
6. 業務(クラスタアプリケーション)	○	○	○*2

異常時の業務継続 ○:可、×:不可

*1 異常検出時は、ノングローバルゾーンを再起動して業務継続可

*2 異常検出時は、業務(クラスタアプリケーション)を再起動して業務継続可

図1.19 Oracle Solarisゾーン環境



監視対象の異常検出方法

1. 物理パーティション

サーバのシステム監視機構と連携した非同期監視が、CPUやメモリ等の異常を契機とするパニック、およびリセットを即時検出し、待機系に切り替えます。

2. 共用ディスク（ディスクアクセスパス）

   ボリューム管理機能（GDS）と組み合わせることで、ディスクアクセスおよび、ディスクアクセスパスの故障を検出（Gdsリソースで監視）し、ディスクアクセス不可または、ディスクアクセスパスの全系故障の場合に待機系に切り替えます。

3. 業務LAN

   ネットワーク多重化機能(GLS)と組み合わせることで、業務LANのネットワークアダプタや経路の故障を検出（Glsリソースで監視）し、ネットワークの全系故障の場合に待機系に切り替えます。

4. OS（グローバルゾーン）

   非同期監視により、OSのパニック、およびリセットを即時検出し、待機系に切り替えます。クラスタインタコネクト(LAN)定周期監視によりOSのハングアップを検出し、待機系に切り替えます。

5. OS(ノングローバルゾーン)

   ー ノングローバルゾーンへの異常(ログイン(zloginコマンド)が不可能)を検出し、待機系に切り替えます。

   ー シングルノードクラスタの場合は、グローバルゾーンのPRIMECLUSTERがノングローバルゾーンを再起動します。

6. 業務（クラスタアプリケーション）

   クラスタアプリケーションのリソース異常発生時に待機系に切り替えます。

   シングルノードクラスタの場合は、ノングローバルゾーンを再起動します。

第2章 PRIMECLUSTERのアーキテクチャ

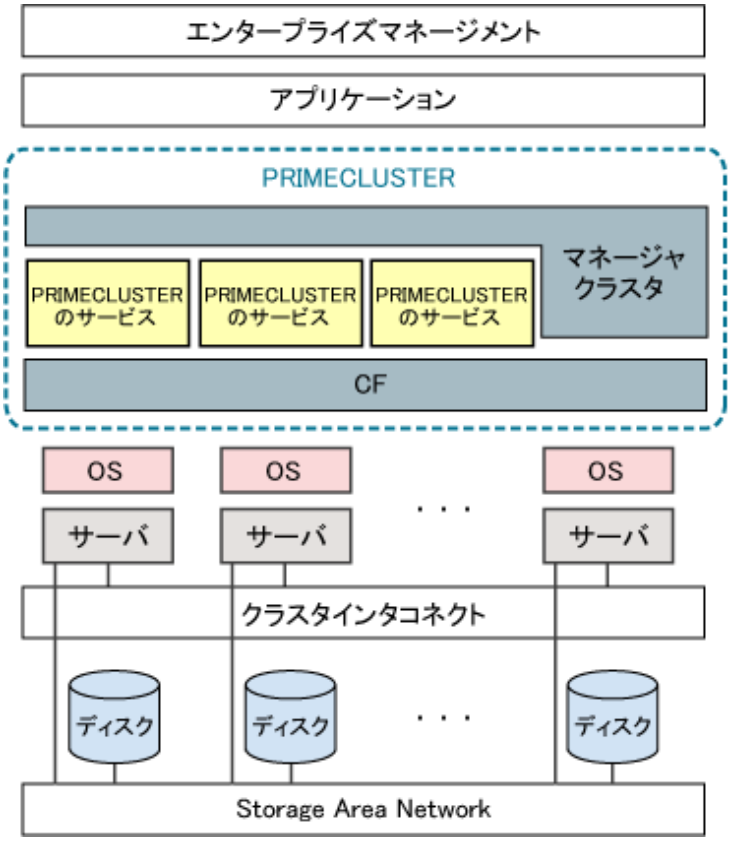
本章では、PRIMECLUSTERのアーキテクチャおよび主な機能について説明します。

2.1 アーキテクチャの概要

PRIMECLUSTERは、高可用性 (HA) を実現するソフトウェアやハードウェアを構築してきた実績に基づいて設計されています。PRIMECLUSTERはこのソリューションとして以下の特徴があります。

- 新しいハードウェアプラットフォーム、オペレーティングシステム、およびクラスタインタコネクトへの容易な移植性
- クラスタシステムを使用/管理するために視覚的、感覚的に理解しやすい操作方法
- 他のアプリケーションがPRIMECLUSTERとの通信機能や、PRIMECLUSTERの機能を利用するためのインタフェースの提供

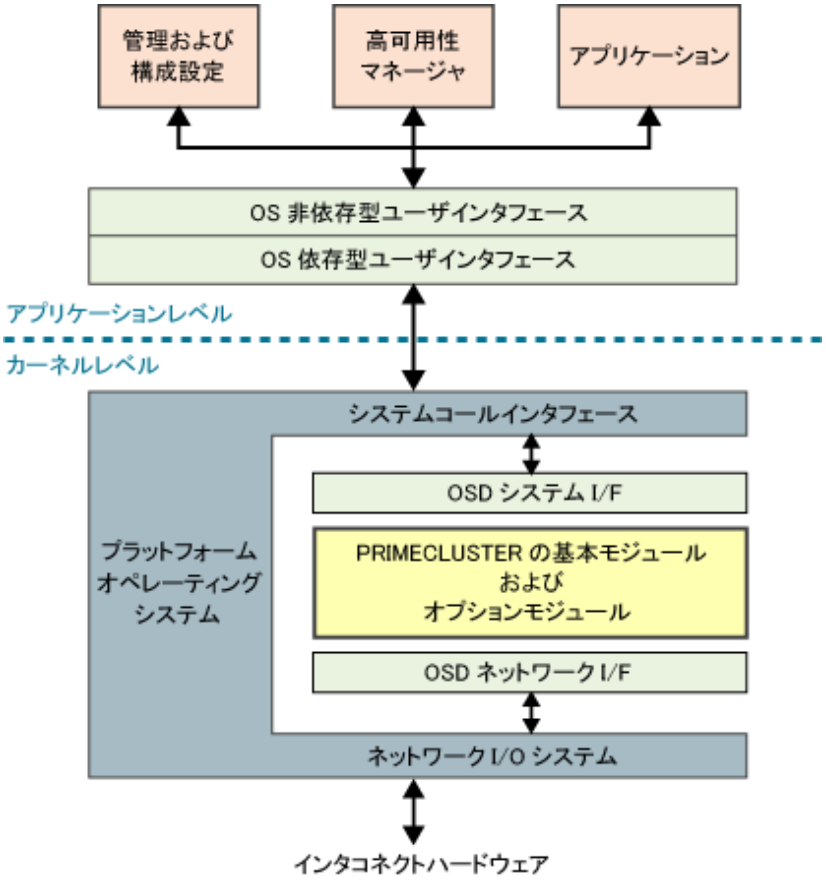
図2.1 一般的なPRIMECLUSTERの設定図



以下の図は、PRIMECLUSTERソフトウェアのアーキテクチャおよびPRIMECLUSTERとオペレーティングシステム本体とのインタフェースの概念を示しています。PRIMECLUSTERの全てのモジュールは、OS依存層 (OS Dependent。以降、OSD層と呼ぶ) に対して、OS非依存型インタフェースを使用して、モジュール間の通信や基本オペレーティングシステムのサービスへのアクセスを行います。OSD層には以下の機能があります。

- メモリアロケーション
- シンクロナイゼーション
- デバイスおよびネットワークアクセス

図2.2 PRIMECLUSTERフレームワークの概要



2.2 PRIMECLUSTER設計理念

PRIMECLUSTERクラスタリングソフトウェアは以下を目標に設計されています。

- モジュール方式
- プラットフォーム非依存性
- 拡張性 (スケーラビリティ)
- 可用性
- データの整合性保証

2.2.1 モジュール方式

PRIMECLUSTERは、基本的なクラスタリング機能を提供する、Cluster Foundation (CF) と呼ばれるモジュールを中心とした集合体で構成されます。PRIMECLUSTERにはParallel Application Services (PAS) モジュール、およびReliant Monitor Services (RMS) モジュールなど、さまざまなモジュールで機能を拡張することが可能です。

2.2.2 プラットフォーム非依存性

PRIMECLUSTER のアーキテクチャはオペレーティングシステムやハードウェアプラットフォームに依存しません。PRIMECLUSTERのモジュールは、オペレーティングシステムのカーネル機構の抽象化に基づいて設計およびコーディングされています。これはオペレーティングシステムやネットワークインタコネクトの種類によって固有のOSD層で処理されます。この方式により、オペレーティングシステムを変更することなしに、PRIMECLUSTERをサポート対象上に実装することができます。これにより、ユーザニーズに合わせたプラットフォームに展開することができます。

2.2.3 拡張性 (スケーラビリティ)

PRIMECLUSTER製品は高可用性に加えて拡張性も備えています。PRIMECLUSTERは、共通のサービスを提供するために、複数のノードを連携して動作させることができます。たとえばPASでは、データベースを複数のノードにわたって並列に動作させることができます。GFSでは、複数ノード上の連携プロセスが同一のデータにアクセスできるようクラスタ規模のファイルシステムを実現しています。

リソース(特にCPU) に対するアプリケーションの要求が、単一マシンの能力を超えてしまっている場合には、PRIMECLUSTERの拡張性が大きな意味を持ちます。ノードをクラスタ化することによって、このようなアプリケーションに対してより大きな処理能力を提供することができます。

2.2.4 可用性

PRIMECLUSTERでは、全てのクラスタ情報をノード間で完全に複製して、ソフトウェアの一点故障を回避します。また、冗長化された複数のクラスタインタコネクトにより、ハードウェアの一点故障 (Single Point of Failure) も回避することができます。また、PRIMECLUSTERのHAマネージャであるRMSにより、ノード障害の発生時に、ユーザ業務をフェイルオーバさせることができるので、業務の可用性を保証します。さらに、PRIMECLUSTERには、ネットワークの可用性を向上させるオプションのネットワーク負荷分散モジュール (GLS) も存在します。

なお、二点以上が同時に故障した場合、データの整合性を保証するため、ユーザ業務のフェイルオーバが発生しない場合があります。

2.2.5 データの整合性保証

PRIMECLUSTERのアルゴリズムは、クラスタパーティション (またはスプリットブレイン状態) 発生時においても、また、複数のハードウェアインタコネクトに障害が発生してもデータの不整合を起こさないように設計されています。クラスタ整合状態 (クォーラム) を利用したアルゴリズムは、クラスタパーティションにより分断されているクラスタシステムの一部のみ動作させます。

2.3 PRIMECLUSTERのモジュール

PRIMECLUSTERのコアコンポーネントであるCluster Foundation (CF) は、全てのコンポーネントの基礎となるクラスタの機能を提供します。CFの構成は以下のとおりです。

- Cluster Admin
  クラスタの構築、管理、運用および診断サービスのインタフェースを提供します。
- Web-Based Admin View
  PRIMECLUSTERの全てのGUIが稼動するフレームワークを提供します。
- クラスタリソース管理機構(Cluster Resource Management)(CRM)

  CRMは各クラスタノード間で同期しているリソースデータベースの管理を行います。クラスタリソースデータベースは、PRIMECLUSTER製品専用のデータベースです。
- PRIMECLUSTER SF
  他のノードを停止させることを保証する機能を提供します。

上記コンポーネントを基盤として、PRIMECLUSTERの機能を強化するオプションコンポーネントを以下に示します。

- Reliant Monitor Services (RMS)
  ユーザ業務の各種プロセスおよび各種リソースの高可用性のため、ユーザ業務のフェイルオーバを制御します。さらに、RMSウィザードによりRMSの容易な設定を可能にします。
- RMSウィザード
  RMS の構成が行えます。
- Parallel Application Services (PAS)
  並列データベースソフトウェアに対する高性能かつ高速な通信機能を提供します。
- GDS
  ディスク装置に格納されているデータの可用性と運用管理性を向上させるボリューム管理機能を提供します。GDSを使用しない場合、ディスクアクセスでエラーが発生した時点ではなく、ディスクアクセスのエラーによってクラスタアプリケーションのリソースが異常になったときに待機系に切り替わるため、待機系への切替えまでに時間がかかります。また、異常箇所の特定にも時間がかかります。
- GFS
  共用ディスク装置に接続している複数ノードによるアクセス機能を備えたファイルシステムを提供します。(Oracle Solaris 10環境のみ)

- GLS
  複数のネットワークインタフェースカード(NIC)を使った冗長化パスを構築することにより信頼性の高い通信機能を可能にします。GLSを導入しない場合、ネットワークの冗長化、VLANの使用、待機系ノードの監視ができません。

2.3.1 CF

CFは、他の全てのPRIMECLUSTERモジュール/コンポーネントが使用するOSD層などの基盤機能を提供します。

CFには以下の特徴があります。

- システム起動時に自動的にロードされる、ロード可能な擬似デバイスドライバの装備
- OSDおよび汎用モジュールを含むCFドライバ

CFはクラスタインタコネクトを使用して、ノードの生存監視、ノード間通信の制御を行います。クラスタインタコネクトの詳細については、“第3章 クラスタインタコネクトの詳細”を参照してください。

2.3.2 Cluster Admin

Cluster Adminは、以下の機能を提供します。

- クラスタシステムの構築
- クラスタシステムの管理
- クラスタシステムの運用および診断機構

Cluster Adminを使って、クラスタシステム内の任意のノードからクラスタシステムの構築、管理、運用を行うことができます。また、ネットワーク経由で、遠隔地のクライアントから管理を行うことも可能です。ユーザはJava対応のWebブラウザを使って管理を行いますが、ノード上でコマンドラインインタフェースを使用することもできます。多様で明解な画面表現やイベントログにより、クラスタの状態に関して簡潔でタイムリーな情報をユーザに提供します。

2.3.3 Web-Based Admin View

Web-Based Admin Viewは、PRIMECLUSTER製品が使用するGUI基盤です。Web-Based Admin Viewの機能を以下に示します。

- 複数のGUIの共通基盤
  PRIMECLUSTER には、CF、RMS、SF を制御する Cluster Admin GUI の他に、GDS や GFS などの他のサービスをサポートする GUI が用意されています。Web-Based Admin View には、これらの全ての GUI が共通基盤として動作します。
- シングルログイン
  1回のログインで複数ノード、複数のGUI製品を使用することが可能です。
- パスワードの暗号化
  クライアントブラウザと管理サーバの間で送信されるパスワードは暗号化されます。
- ロギング
  構成設定または管理に関するすべてのGUI操作をロギングします。
- 3層構造
  管理サーバをクラスタシステムと分離した外部に設定します。

参照

Web-Based Admin Viewの機能の詳細については、“PRIMECLUSTER Web-Based Admin View操作手引書” を参照してください。

2.3.4 クラスタリソース管理機構(Cluster Resource Management)(CRM)

CRMは各クラスタノード間で同期しているリソースデータベースの管理を行います。クラスタリソースデータベースは、PRIMECLUSTER製品専用のデータベースです。他のアプリケーションに使用できる汎用のデータベースではありません。

CRMは、CIPを使用してノード間でリソースデータベースの一致化を行い、PRIMECLUSTER内のコンポーネントで使用しているリソースデータベースが、すべてのノードで同一となるよう管理します。

2.3.5 PRIMECLUSTER SF

PRIMECLUSTERシャットダウン機構 (PRIMECLUSTER SF) は、クラスタシステム内でユーザ資産に対する競合が発生するような異常処理時に、他のノードを停止させることを保証する機能を提供します。PRIMECLUSTER SF は主に以下のコンポーネントで構成されます。

- SD (シャットダウンデーモン)
  クラスタノードの状態を監視し、状態を収集したり、ノードの手動シャットダウンを要求したりするためのインタフェースを提供します。
- SA (シャットダウンエージェント)
  リモートクラスタノードを停止させることを保証します。
- MA (非同期監視)
  SAの機能に加え、リモートクラスタノードの状態を監視し、そのノードのダウンを即時に検出します。
  クラスタノードを強制停止させる経路を定期的 (10分間隔) に確認します。

SA (シャットダウンエージェント)

シャットダウンエージェントはリモートクラスタノードの確実な停止を保証します。シャットダウンエージェントは、クラスタノードのアーキテクチャによって異なります。

シャットダウンエージェントは以下の機能を提供します。

- ノードの強制停止

  異常が発生したノードの強制停止を保証します。

- オプションハードウェアの接続確認（シャットダウンエージェントのテスト）

  ノードの強制停止で使用するオプションハードウェアへの接続が正しく行えるかを定期的 (10分間隔) に確認します。

PRIMECLUSTER SFでは、以下のシャットダウンエージェントを提供します。

- RCI (SA_pprcip, SA_pprcir)
  Remote Cabinet Interface

  SPARC Enterprise Mシリーズに搭載されるハードウェアの1つ、RCIを利用して、他ノードを意図的にパニックまたはリセットさせることで、確実なノード停止を実現します。

- XSCF (SA_xscfp, SA_xscfr, SA_rccu, SA_rccux)
  eXtended System Control Facility

  SPARC Enterprise Mシリーズに搭載されるハードウェアの1つ、XSCFを利用して、他ノードを意図的にパニックまたはリセットさせることで、確実なノード停止を実現します。
  また、コンソールにXSCFを使用している場合は、他ノードにbreak信号を送信して確実なノード停止を実現します。

- XSCF SNMP (SA_xscfsnmpg0p, SA_xscfsnmpg1p, SA_xscfsnmpg0r, SA_xscfsnmpg1r, SA_xscfsnmp0r, SA_xscfsnmp1r)
  eXtended System Control Facility Simple Network Management Protocol

  SPARC M10、M12 に搭載されるハードウェアの1つ、XSCF を利用して、他ノードを意図的にパニックまたはリセットさせることで、確実なノード停止を実現します。

- ALOM (SA_sunF)
  Advanced Lights Out Management

  SPARC Enterprise T1000、T2000 の ALOM を利用して、他ノードに break 信号を送信して確実なノード停止を実現します。

- ILOM (SA_ilomp, SA_ilomr)
  Integrated Lights Out Manager

  SPARC Enterprise T5120、T5220、T5140、T5240、T5440、SPARC T3、T4、T5、T7、S7シリーズのILOM を利用して、他ノードを意図的にパニックまたはリセットさせることで、確実なノード停止を実現します。

- KZONE(SA_kzonep, SA_kzoner, SA_kzchkhost)

  Oracle Solaris カーネルゾーン

  SPARC M10、M12、SPARC T4、T5、T7、S7シリーズでOracle Solaris カーネルゾーンを使用している場合、他ノード（カーネルゾーン）を意図的にパニックまたはリセットさせることで、確実なノード停止を実現します。

また、グローバルゾーンホストの状態を確認し、グローバルゾーンホストが停止した場合に他ノード(カーネルゾーン)が停止状態であると判断します。グローバルゾーンホストの強制停止は行いません。

- BLADE (SA_blade)

  PRIMERGYブレードサーバで使用可能な機能で、SNMPコマンドを使用して、他ノードをシャットダウンさせることで、確実なノード停止を実現します。

- IPMI (SA_ipmi)
  Intelligent Platform Management Interface

  PRIMERGYに搭載されるハードウェアの1つであるBMC(Baseboard Management Controller)、またはiRMC(integrated Remote Management Controller)をIPMIで操作して、他ノードをシャットダウンさせることで、確実なノード停止を実現します。

- kdump (SA_lkcd)

  PRIMERGY 、PRIMERGY ブレードサーバで kdump を使用して、他ノードをパニックさせることで、確実なノード停止を実現します。

- MMB (SA_mmbp, SA_mmbr)
  Management Board

  PRIMEQUEST 2000に搭載されるハードウェアの1つ、MMBを利用して、他ノードを意図的にパニックまたはリセットさせることで、確実なノード停止を実現します。

- iRMC (SA_irmcp, SA_irmcr, SA_irmcf)

  PRIMEQUEST 3000に搭載されるハードウェアのiRMC/MMBを利用して、他ノードを意図的にパニック、リセット、または電源切断させることで、確実なノード停止を実現します。

注意

PRIMERGYのiRMCでは本シャットダウンエージェントは使用できません。

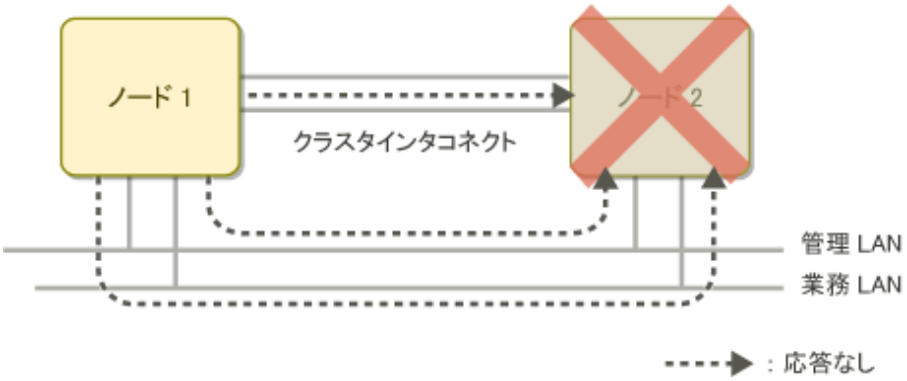
- ICMP (SA_icmp)

  ネットワーク経路を使用して他ノードの状態を確認し、他ノードから応答がない場合に停止状態であると判断します。
  他ノードの強制停止は行いません。

  以下の図は、2ノードのクラスタシステムにおいて、1つのノード(ノード2)が停止した場合のSA_icmp による状態確認の例です。

  指定されたすべてのネットワーク経路でノード2から応答がなかった場合、SA_icmpはノード2が停止状態であると判断します。

図2.3 他ノードが停止している場合の SA_icmp による状態確認

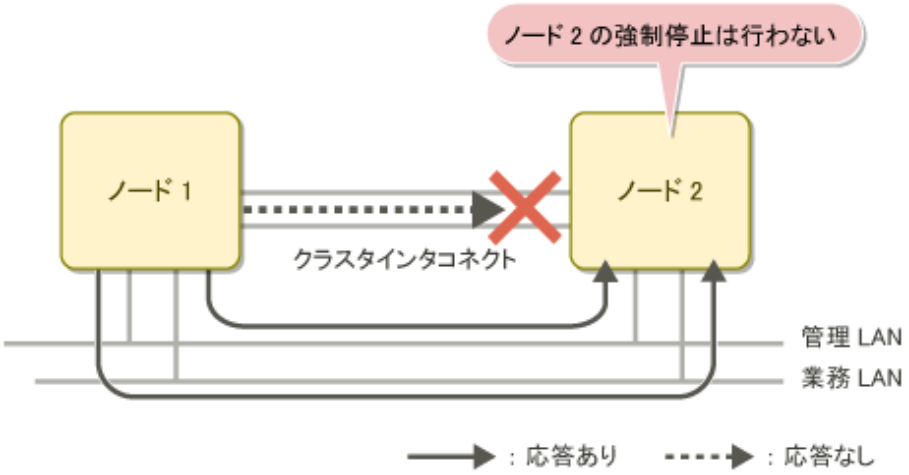


以下の図は、2ノードのクラスタシステムにおいて、クラスタインタコネクトが故障した場合のSA_icmp による状態確認の例です。

指定されたネットワーク経路のいずれかで、ノード2からノード1に応答があった場合、SA_icmp はノード2が運用状態であると判断します。

この場合、SA_icmp はノード2の強制停止を行いません。

図2.4 クラスタインタコネクトが故障している場合の SA_icmp による状態確認



- VMCHKHOST (SA_vmchkhost)

  KVM 仮想マシン機能で管理OSにクラスタシステムを導入している場合、管理OSのクラスタシステムと連携してゲストOSの状態を確認します。
  他ノードの強制停止は行いません。

- libvirt (SA_libvirtgp, SA_libvirtgr)

  PRIMERGY 、PRIMERGY ブレードサーバ 、PRIMEQUEST 3000/2000 シリーズで KVM 仮想マシン機能を使用している場合、他ノード（ゲストOS）を意図的にパニックまたはリセットさせることで、確実なノード停止を実現します。

- VMware vCenter Server連携(SA_vwvmr)

  VMware vCenter Serverと連携し、他ノード(ゲストOS)を意図的に電源断させることで、確実なノード停止を実現します。

- FUJITSU Cloud Service K5 API(SA_vmk5r)

  FUJITSU Cloud Service K5 APIを利用して、他ノード(インスタンス)を意図的にシャットダウンまたは電源断させることで、確実なノード停止を実現します。

- OpenStack API(SA_vmosr)

  OpenStack APIを利用して、他ノード(インスタンス)を意図的に再起動させることで、確実なノード停止を実現します。

MA(非同期監視)

非同期監視は、ハードウェア特性を活かしてノードの状態を監視し、ノードダウンを即時に検出します。PRIMECLUSTERシステムは、クラスタインタコネクトを利用した、ハートビートの送信と応答によるノードの状態監視を定周期間隔で行っていますが、非同期監視を利用することにより、より即時的なノードのダウン検出を実現します。

非同期監視は以下の機能を提供します。

- ノードの状態監視

  非同期監視は、ハードウェアが提供する機能を利用したノードの状態監視を行います。突然のシステムパニックや電源切断など、万が一他のノードに異常が発生した場合、SFにその異常を通知します。また、システム負荷 (System Load) が著しく高いことが原因で、クラスタノード間でのハートビート要求の送信と応答が一時的に途切れた場合でも、非同期監視がオプションハードウェアを経由してノードの状態を正確に判断します。

- ノードの強制停止

  SA (シャットダウンエージェント) としての機能を提供し、異常が発生したノードの強制停止を保証します。

- オプションハードウェアの接続確認（シャットダウンエージェントのテスト）

  SA (シャットダウンエージェント) としての機能を提供し、ノードの状態監視やノードの強制停止で使用するオプションハードウェアへの接続が正しく行えるかを定期的 (10分間隔) に確認します。

PRIMECLUSTER SFでは、以下の非同期監視を提供します。

RCI非同期監視 (SPARC Enterprise Mシリーズ)

SPARC Enterprise Mシリーズに搭載されるハードウェアの1つ、RCIを利用してノードの状態を監視する機能です。ハードウェア本体に標準で実装されているシステム監視機構(System Control Facility: SCFと略する)がハードウェアの状態を監視し、その状態をソフトウェアに通知することでノードダウンを判断することができます。
また、他ノードを意図的にパニックまたはリセットさせることで確実な強制停止を実現し、ユーザ資産への競合を防ぎます。

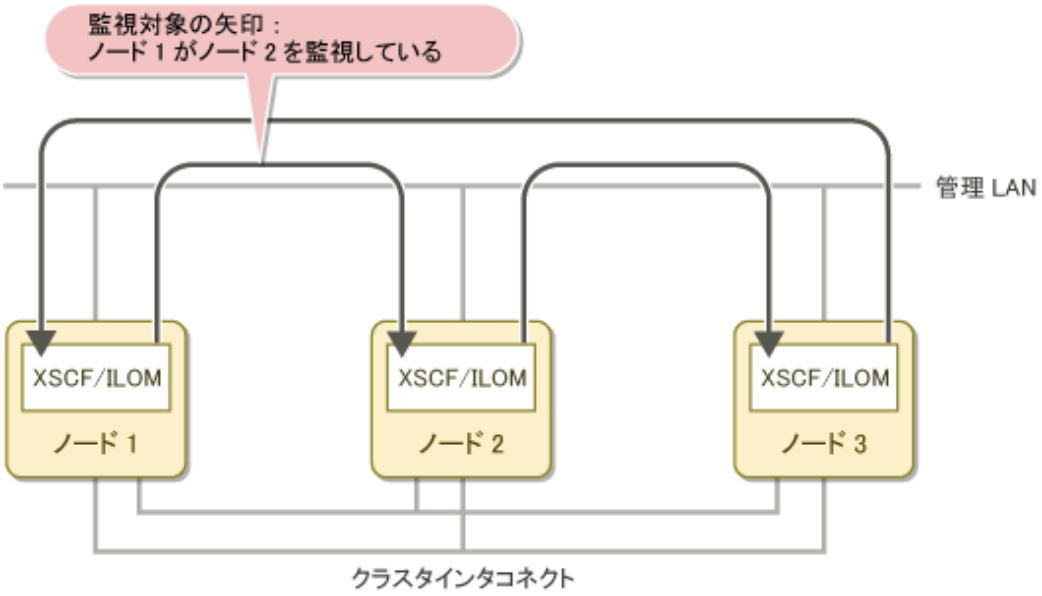
コンソール非同期監視

XSCF/ILOMを使用して、クラスタシステムを構成する各ノードのコンソールに表示されるメッセージを監視する機能です。パニック発生時等のコンソールメッセージを他ノードが検出し、メッセージ出力ノードのノードダウンを判断します。コンソール非同期監視は通常、数珠状に1対1の関係で他ノードの状態を監視しており、異常が発生してノードがダウンした場合、ダウンノードが監視していたノードを監視する役目を、その他のノードに引き継ぎます。また、ノードに対してbreak 信号を送信して確実なノード停止を行います。

ノードがダウンした場合の監視の引き継ぎについて、3 ノードで構成されるクラスタシステムを例に示します。

以下の図は、3ノードのクラスタシステムにおいて、1つのノードが停止した場合に監視機能がどのように引き継がれるかを示しています。矢印は、どのノードがどのノードを監視しているかを表します。

図2.5 正常稼動時のコンソール非同期監視の処理



ノード2に異常が発生してダウンすると、以下の処理が実行されます。

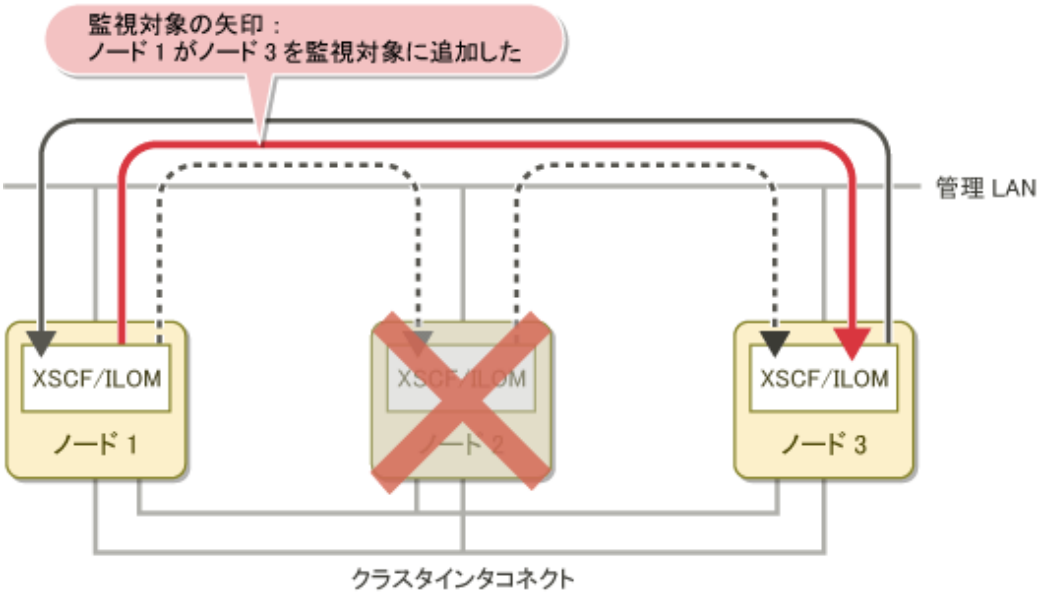
- ノード1はノード3の状態監視を開始します。
- 以下のメッセージがノード1の/var/adm/messagesファイルに出力されます。

  FJSVcluster: INFO: DEV: 3044: The console monitoring agent took over monitoring (node: *targetnode*)

  FJSVcluster: 情報: DEV: 3044: コンソール非同期監視機能の監視対象にノード*targetnode*を追加しました。

以下の図は、ノード2が停止した場合に、ノード1がノード3を監視対象ノードとして追加する様子を示しています。

図2.6 ノード異常発生時のコンソール非同期監視の処理



注意

コンソール非同期監視が停止していたときに監視機能の引き継ぎが行われた場合、停止していたコンソール非同期監視が再開されます。

ノード2が異常から復旧後に起動してくると、以下の処理が実行されます。

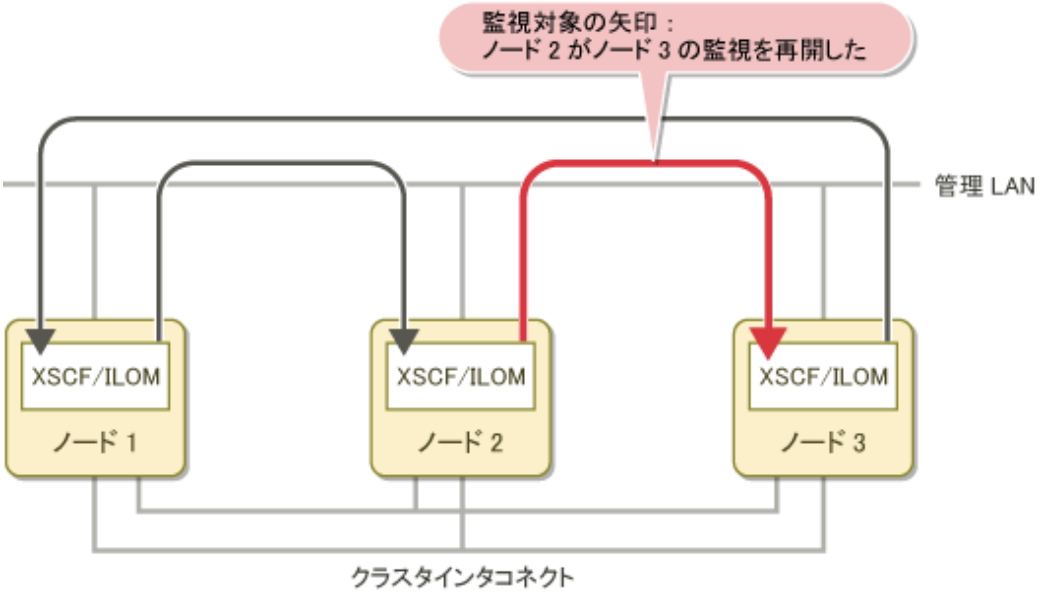
- 従来の正常起動時の監視形態に戻ります。
- 以下のメッセージがノード1の/var/adm/messages ファイルに出力されます。

  FJSVcluster: INFO: DEV: 3045: The console monitoring agent cancelled to monitor (node: *targetnode*)

  FJSVcluster: 情報: DEV: 3045: コンソール非同期監視機能の監視対象からノード*targetnode*を削除しました。

以下の図は、クラスタに復旧したノード2がノード3の監視を再開する様子を示しています。

図2.7 ノード復旧時のコンソール非同期監視の処理



注意

コンソール非同期監視では、コンソールのメッセージを監視しているため、突然の電源切断の状態を判断できずLEFTCLUSTER状態が発生します。本現象が発生した場合は、ノードにDOWNマークを付ける必要があります。DOWNマークの付けかたについては、“PRIMECLUSTER Cluster Foundation 導入運用手引書”を参照してください。

SNMP非同期監視 (SPARC M10、M12)

SPARC M10、M12 に搭載されるシステム監視機構(eXtended System Control Facility: XSCFと略する)を利用して、ノードの状態を監視する機能です。
XSCFがハードウェアの状態を監視し、その状態をSNMP(Simple Network Management Protocol)を使用して、ソフトウェアに通知することで、ノードダウンを判断することができます。
また、他ノードを意図的にパニックまたはリセットさせることで、確実な強制停止を実現し、ユーザ資産への競合を防ぎます。

MMB非同期監視 (PRIMEQUEST 2000)

PRIMEQUEST 2000に搭載されるハードウェアの1つ、MMBを利用してノードの状態を監視する機能です。ハードウェア本体に標準で実装されているMMBがハードウェアの状態を監視し、その状態をソフトウェアに通知することでノードダウンを判断することができます。
また、他ノードを意図的にパニックまたはリセットさせることで確実な強制停止を実現し、ユーザ資産への競合を防ぎます。

iRMC非同期監視 (PRIMEQUEST 3000)

PRIMEQUEST 3000に搭載されるハードウェアのiRMCとMMBを利用して、ノードの状態を監視する機能です。ハードウェア本体に標準で実装されているiRMCとMMBがハードウェアの状態を監視し、その状態をソフトウェアに通知することでノードダウンを判断することができます。

また、他ノードを意図的にパニック、リセット、または電源切断させることで確実な強制停止を実現し、ユーザ資産への競合を防ぎます。

注意

PRIMERGYのiRMCでは本非同期監視は使用できません。

注意

RCI 非同期監視のノード状態の監視は、/var/adm/messages ファイルに以下のメッセージ(a) が出力されてから、(b) が出力されるまでの間機能しています。
コンソール非同期監視の場合は、それぞれ(c) と(d) に該当します。
SNMP非同期監視の場合は、それぞれ(e) と(f) に該当します。
MMB非同期監視の場合は、それぞれ (g) と (h) に該当します。
iRMC非同期監視の場合は、それぞれ (i) と (j) に該当します。
ノード状態の監視が機能していない状態では、ノードを強制的に停止する機能が正常に動作しないことがあります。

(a) FJSVcluster: INFO: DEV: 3042: The RCI monitoring agent has been started

FJSVcluster: 情報: DEV: 3042: RCI非同期監視機能を開始しました。

(b) FJSVcluster: INFO: DEV: 3043: The RCI monitoring agent has been stopped

FJSVcluster: 情報: DEV: 3043: RCI非同期監視機能を停止しました。

(c) FJSVcluster: INFO: DEV: 3040: The console monitoring agent has been started (node:*monitored node name*)

FJSVcluster: 情報: DEV: 3040: コンソール非同期監視機能を開始しました。
(node:監視対象ノード名)

(d) FJSVcluster: INFO: DEV: 3041: The console monitoring agent has been stopped (node:*monitored node name*)

FJSVcluster: 情報: DEV: 3041: コンソール非同期監視機能を停止しました。
(node:監視対象ノード名)

(e) FJSVcluster: INFO: DEV: 3110: The SNMP monitoring agent has been started.

FJSVcluster: 情報: DEV: 3110: SNMP 非同期監視を開始しました。

(f) FJSVcluster: INFO: DEV: 3111: The SNMP monitoring agent has been stopped.

FJSVcluster: 情報: DEV: 3111: SNMP 非同期監視を停止しました。

(g) FJSVcluster: INFO: DEV: 3080: The MMB monitoring agent has been started.

FJSVcluster: 情報: DEV: 3080: MMB 非同期監視を開始しました。

(h) FJSVcluster: INFO: DEV: 3081: The MMB monitoring agent has been stopped.

FJSVcluster: 情報: DEV: 3081: MMB 非同期監視を停止しました。

(i) FJSVcluster: INFO: DEV: 3120: The iRMC asynchronous monitoring agent has been started.

FJSVcluster: 情報: DEV: 3120: iRMC 非同期監視を開始しました。

(j) FJSVcluster: INFO: DEV: 3121: The iRMC asynchronous monitoring agent has been stopped.

FJSVcluster: 情報: DEV: 3121: iRMC 非同期監視を停止しました。

2.3.6 RMS

RMSは、2ノード以上の構成でクラスタのハードウェアおよびソフトウェアの可用性を保証する、HAマネージャです。PRIMECLUSTERは、各コンポーネントの冗長化と、稼動中のノードへの監視対象リソースのフェイルオーバ機能により可用性を保証します。

たとえば、以下に示すようなシステムコンポーネントが監視対象リソースになります。

- ファイルシステム
- ボリューム (ディスク)
- アプリケーション
- ネットワークインタフェース
- ノード全体

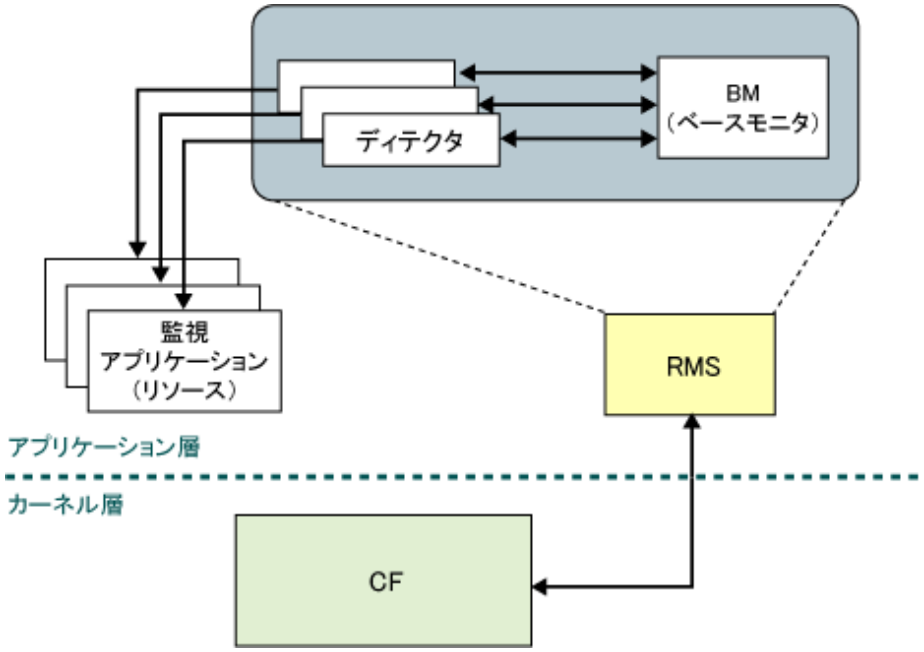
RMSでは、クラスタシステムの複数ノードを使用し、各ノードは他のノードのリソースを引き継ぐように設定され、ユーザ業務は冗長化されます。

ユーザ業務の可用性は、ディテクタプログラムを使ったリソース監視により保証されます。リソースに障害が発生すると、RMSはユーザ定義のリカバリ処理を起動します。このリカバリ処理が、他のノード上でリソースを使用できるようにするトリガになります。

依存関係にある複数リソースは、グループ化することで、グループ内のリソースの一部に障害が発生したときにグループ全体に対するリカバリ処理を実行されるようにすることも可能です。フェイルオーバ時には、元のノード上の全てのリソースを確実にオフラインにしてから、新しいノード上でリソースをオンラインにします。これにより、複数ノードが同時に1つのリソースにアクセスしようとする競合により、データが破損する可能性を排除します。

以下の図は、RMSのディテクタによるリソース監視方法を示しています。ディテクタからRMS BM (ベースモニタ) にリソースの状態変化が通知されると、RMSは対処が必要であるかどうか判断します。

図2.8 RMSのリソース監視



2.3.6.1 RMSウィザード

RMSウィザードは、RMSが動作するための構成を作成する機能を提供します。

- RMS Wizard Tools
  RMSウィザードの基盤および、RMS BM (ベースモニタ) とのインタフェースを提供します。RMS Wizard Toolsにより、RMS構成を作成する設定が簡素化され、RMSとの統合によりHAクラスタ環境が強化されます。

2.3.6.2 プロセス監視機構

プロセス監視機構はRMSに対してプロセスの状態を通知します。プロセス監視機構の長所は以下のとおりです。

- プロセスの状態を素早く低い負荷によりRMSに通知します。これはユーザアプリケーションの高速な切替えにつながります。
- 不慮のエラーにより終了したプロセスを自動的に再起動します。
- プロセスの状態を確認するためのコマンドを用意する必要がありません。

2.3.7 PAS

PAS (Parallel Application Services) は、OPS (Oracle Parallel Server) のような並列データベースアプリケーションに対する高性能かつ高速な通信機能を提供します。

並列データベースアプリケーションはクラスタテクノロジーを採用した市販アプリケーションです。複数ノードで構成されているクラスタシステムの、各ノード間で負荷およびデータを分割することにより、並列データベースアプリケーションは1台の大規模SMPサーバの限界を超えるパフォーマンスを実現します。

2.3.8 GDS

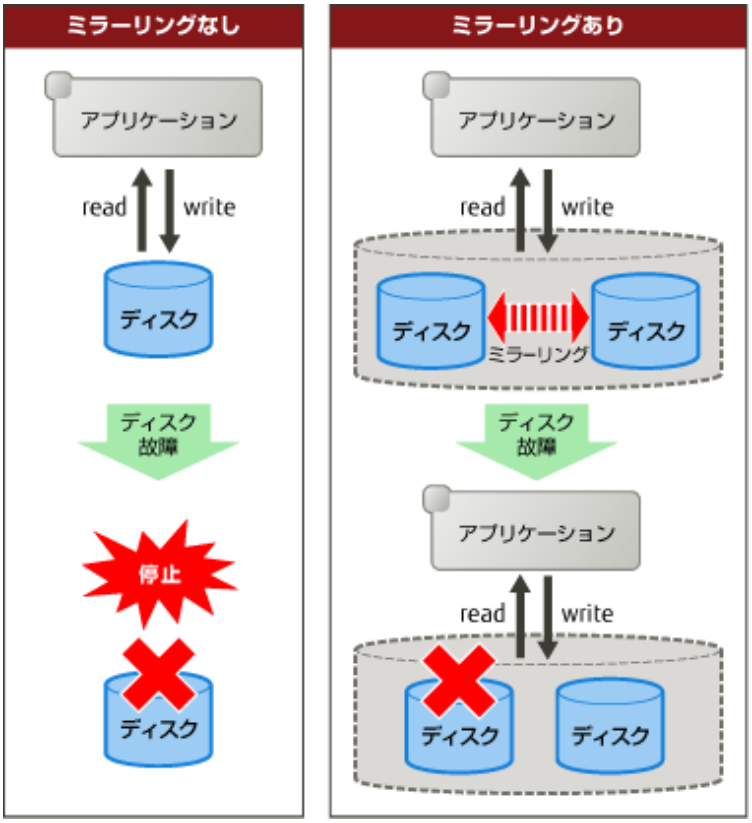
GDS は、ディスク装置に格納されているデータの可用性と運用管理性を向上させるボリューム管理ソフトウェアです。GDSは、ハードウェアの故障やユーザの操作ミスからデータを保護し、ディスク装置の運用管理を支援します。

ボリューム管理機能には以下の2つの役割があり、それらは密接に関連しています。

- ディスクデータの可用性の向上
- ディスクデータの運用管理性の向上

GDSのミラーリング機能は、ディスクデータの複製を複数のディスクに保持することにより、ハードウェアの故障からデータを保護します。これにより、不測のトラブルが発生しても、ユーザはアプリケーションを停止することなくディスクデータへのアクセスを継続できます。

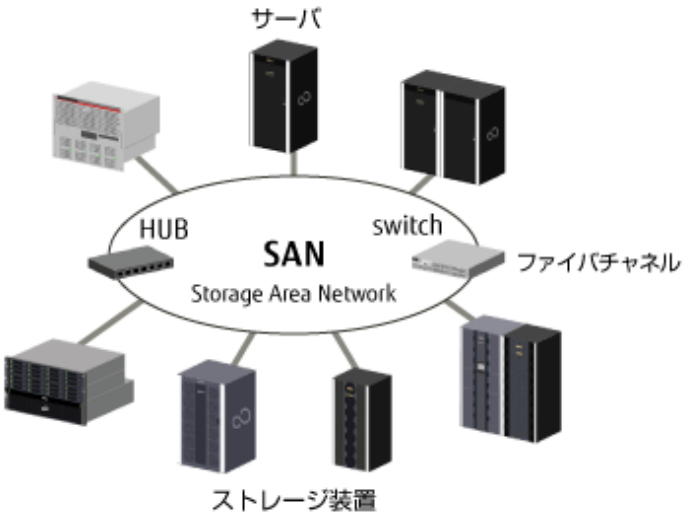
図2.9 ディスクミラーリング



また、GDSの運用管理機能は、さまざまなディスク管理においてシステム管理者の負担を軽減します。使いやすい運用管理機能には、管理作業を簡易にするだけではなく、操作ミスによるデータ破壊を防止する効果があります。

SAN (Storage Area Network)においては、複数のサーバと複数のディスク装置が自由に接続されるため、ディスク装置のデータを複数のサーバから直接共用することができます(下図を参照)。これにより、ファイルシステムやデータベースの同時共用が可能になります。また、サーバ間でのデータの複写や、バックアップなどの作業の利便性が改善されます。その反面、複数のサーバからのアクセス競合によってデータ破壊が発生するという問題が潜在しているため、SANに適合したボリューム管理機能が不可欠です。

図2.10 SAN (Storage Area Network)



GDSは、SANに適合したボリューム管理機能を提供します。GDSを使用することにより、特定のサーバにローカル接続されたディスク装置だけでなく、SANを経由して複数のサーバに共用接続されたディスク装置も含めて、全てのサーバに接続された、全てのディスク装置を統一的に管理することができます。

注意

PRIMERGY(VMware環境は除く)では、システムディスクは管理できません。

GDSの主な機能を以下に示します。

- システムディスクのミラーリング機能 (Solarisサーバ、PRIMEQUEST、およびVMware環境)
- 共用ディスクのミラーリング機能
- ディスクアレイ筐体間のミラーリング機能
- 運用ノードと待機ノードのローカルディスクをネットワーク経由でミラーリングするサーバ間ミラーリング機能 (Linuxサーバ)
- ディスク故障時に自動的にミラーリング状態を回復するホットスペア機能
- アプリケーションを停止することなく、故障したディスクを交換するホットスワップ機能
- システムダウン後やクラスタフェイルオーバ後にミラーリング状態を高速に回復する高速等価性回復機能
- SAN環境に接続されたディスクの統合管理とアクセス制御
- サーバとディスク装置との物理的な接続構成を自動的に認識し、構成情報の登録や接続構成のチェックを行う自動構成機能 (Solarisサーバ)
- 大容量ボリュームの作成を可能にするコンカチネーション機能
- ディスクへのアクセス負荷を分散するストライピング機能
- 柔軟なディスク構成を実現する論理パーティション分割機能
- 主業務への影響を最小限に抑えたバックアップ運用を支援するスナップショット機能

参照

詳細については、“PRIMECLUSTER Global Disk Services説明書”を参照してください。

2.3.9 GFS

GFS は、複数のノードから同時共用できる GFS 共用ファイルシステムを提供します。

Solaris版はSolaris10で利用可能です。

2.3.9.1 GFS共用ファイルシステム

GFS共用ファイルシステムは共用ディスク装置に接続している複数ノード (最大2ノード)による同時アクセス機能を備えた信頼性と性能に優れたファイルシステムです。
GFS共用ファイルシステムには主要な機能に加え、以下のような機能があります。

- 複数ノードから1つのファイルまたはファイルシステムに同時にアクセスする機能
- 複数ノードからファイルデータを参照または更新する場合に整合性を維持する機能
- ノードが停止中のとき、他のノードでファイル操作を続行し、ファイルシステムの整合性を維持する機能
- ファイルシステムの高速修復機能
- 共用記憶装置への直接アクセスによりデータアクセス性能を向上させる機能
- ファイル領域の連続したブロックを割当てることによる高速I/O処理機能
- 各ノードのファイルキャッシュを利用したファイルアクセス機能
- マルチボリュームのサポートによるI/Oの負荷分散および大容量記憶ファイル
- ファイルシステムを再構築せずに拡張する機能

- WebブラウザでGUIによりファイルシステムを操作する機能

整合性が保証された同時共用アクセス機能

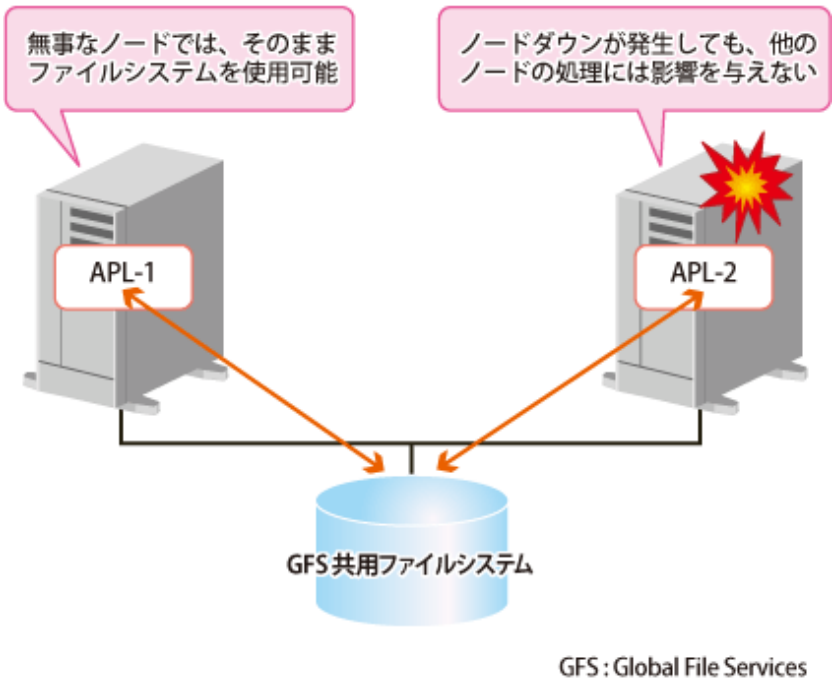
GFS共用ファイルシステムでは、複数ノードからデータを更新する場合に整合性が保証されます。
また、ノードをまたいだファイルロック機能も従来のUNIXファイルシステムのアプリケーションプログラムインタフェース (API) で利用することができます。このように分散アプリケーションが複数ノードで実行される場合、従来のUNIXファイルシステムのAPIをそのまま使って適切なアプリケーションデータを転送することができます。

高可用性(HA)

GFS共用ファイルシステムを複数ノードで使用する場合、1つのノードが停止中でも他のノードからファイルにアクセスすることが可能です。停止中のノードが保持しているファイルシステム情報の整合性は、他のノードのGFS共用ファイルシステムで自動的に回復されます。そのため、他のノードで実行中のアプリケーションは、ファイルシステムの操作エラーを起こすことなく処理を続行します。

ファイルの作成や削除など、ファイルシステム構造の変更に必要な処理はGFS共用ファイルシステムの更新ログと呼ばれる領域に記録されます。この領域に保存された情報を使うことにより、ファイルシステム構造全体をチェックしなくてもシステム障害をリカバリすることができます。

図2.11 ノードが停止中のときの処理続行



データアクセス性能

GFS共用ファイルシステムでは、共用ディスク装置のファイルシステムに複数ノードからアクセスすることができます。従来の分散ファイルシステムでは、データはLANによるネットワーク通信により、ファイルシステムデータを管理するサーバからのアクセスを要求したクライアントに転送されます。これに対してGFS共用ファイルシステムでは、要求側ノードからディスク装置に直接アクセスします。そのため、NFSよりネットワーク負荷が軽く、要求の読込み/書込みの応答時間も速くなります。

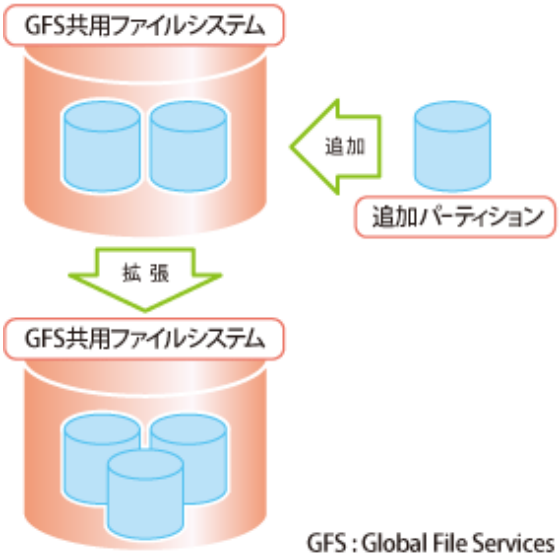
また、GFS共用ファイルシステムではファイルデータに連続したブロックを割当てることによりI/O処理をまとめて行い、ファイルシステムの性能を改善します。

GFS共用ファイルシステムには複数のパーティションを1つのファイルシステムに統合する機能があります。複数パーティション構成の場合、ラウンドロビン (Round Robin) 割当て方式を使用してファイルごとに異なるパーティションのファイルデータ領域を使用するようにします。そのため、I/O負荷は複数ディスク装置に分散され、ファイルシステムの性能が向上します。この機能を使用することにより後述するデータパーティションの追加が容易になります。

拡張性 (スケーラビリティ)

GFS共用ファイルシステムでは、空のディスクパーティションを指定するだけでファイルシステムを簡単に拡張することができます。この機能により、ファイルシステムの空き領域不足がすぐに解消されます。

図2.12 パーティションの追加



2.3.9.2 メリット

GFS には以下のメリットがあります。

- LANを経由しない直接アクセスなどの優れたファイルアクセス機能や各システムのファイルキャッシュを使用することにより、システム全体の性能が向上します。
- 複数のシステム上のファイルによりリンクされたアプリケーションを分散実行することによりCPU の負荷分散が実現し、データの整合性も保証されます。
- 1つのノードが停止しても他のノード上でファイルアクセスを続行することのできる可用性の高いシステムを実現します。
- エクステントベースの領域管理、マルチボリューム機能などによる高速ファイルアクセスを実現します。
- 複数ディスク装置をマルチボリューム機能や領域不足の場合の領域拡張などの機能と連携させることにより、大規模なファイルシステムの作成などのファイルシステムのリソース管理が容易になります。
- ファイルシステムをWeb ブラウザで操作することができるため、環境の構築や管理が容易になります。

2.3.10 GLS

GLS は、複数のNICを使用して、自システムが接続されるネットワーク伝送路を冗長化し、通信全体の高信頼化を実現するソフトウェアです。

GLSには以下のメリットがあります。

- 複数のNICにより伝送路を冗長化することで、耐故障性や可用性に優れた信頼性の高いネットワークを構築することができます。
- GLSを利用するアプリケーションは、冗長化した伝送路の構成や伝送路上で発生したネットワーク障害を意識することなく、業務を行うことができます。

GLS には次の2つの機能があります。

- 伝送路二重化機能
  - ー 高速切替方式
    同一ネットワーク上のサーバ間の伝送路を多重化
  - ー NIC切替方式
    同一ネットワーク上のサーバとスイッチ/HUB間の伝送路を二重化

- 仮想NIC方式(Solaris)
  同一ネットワーク上のサーバとスイッチ/HUB間の伝送路を二重化。二重化した経路を仮想環境(ゲストドメインやノングローバルゾーン)で使用することで、効率的に仮想サーバ集約を実現。
- 仮想NIC方式(Linux)
  同一ネットワーク上のサーバとスイッチ/HUB間の伝送路を二重化。二重化した経路を仮想環境(KVM仮想マシン機能のゲストOS)で使用することで、効率的に仮想サーバ集約を実現。
- GS/SURE連携方式 (Solaris)、GS連携方式(Linux)
  同一ネットワーク上のサーバ、グローバルサーバ/SURE SYSTEM、およびExINCA間の伝送路を多重化

・ マルチパス機能

- マルチパス方式(Solaris)
  同一ネットワーク上のサーバとスイッチ間の伝送路を多重化
- マルチリンクイーサネット方式(Solaris)
  同一ネットワーク上のサーバとスイッチ間で多重化した伝送路の 送信データ負荷分散

参照

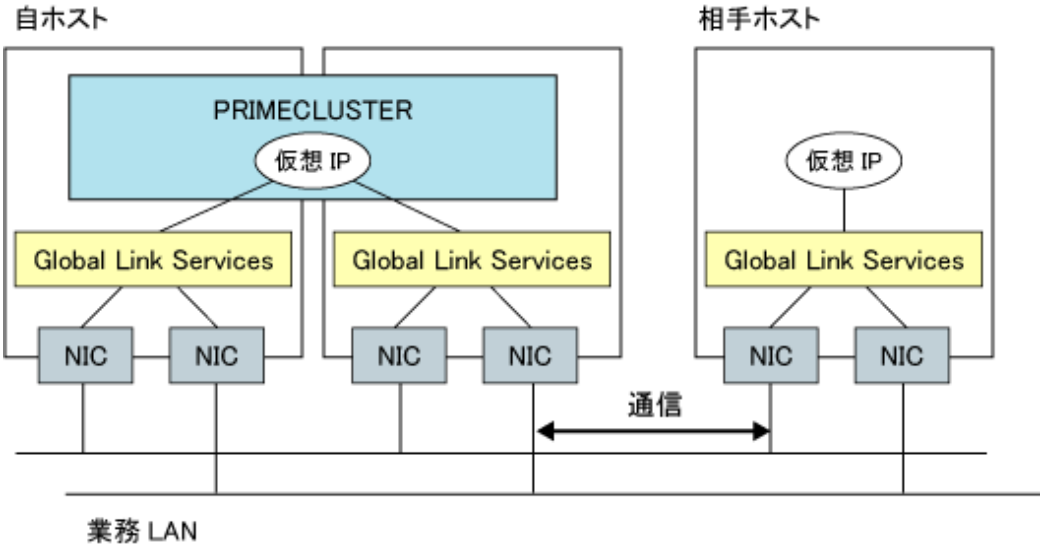
GLSの機能の詳細については、“PRIMECLUSTER Global Link Services 説明書 (伝送路二重化機能編)” 、“PRIMECLUSTER Global Link Services 説明書 (伝送路二重化機能 仮想NIC方式編)”および “PRIMECLUSTER Global Link Services 説明書 (マルチパス機能編)” を参照してください。

なお、伝送路二重化機能 仮想NIC方式編およびマルチパス機能編はSolaris版のみ存在します。

2.3.10.1 高速切替方式

同一ネットワーク上のSolarisまたはLinuxサーバ間の伝送路を冗長化し、伝送路障害発生時の通信継続、および伝送路同時使用によるトータルスループットの向上を実現します。本方式では、冗長化した伝送路を同時に使用し、障害発生時は該当の伝送路を切り離して縮退運用します。GLS自身が制御するため、障害を早期に検出することが可能です。通信可能な相手装置は、SPARC Servers、PRIMEPOWER、GP7000F、富士通S series、GP-S、PRIMERGY、およびPRIMEQUESTです。なお、ルータを超えた別ネットワーク上のホストとの通信には利用できません

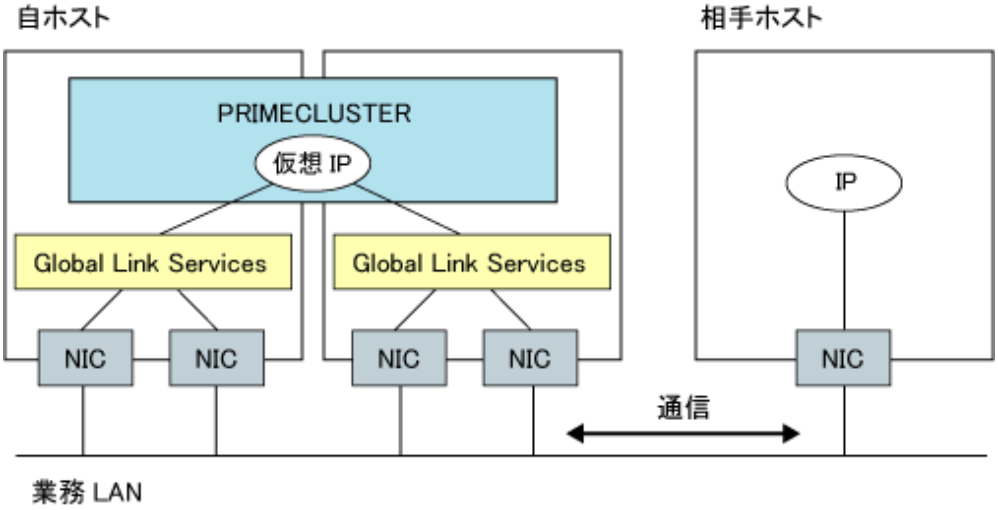
図2.13 高速切替方式



2.3.10.2 NIC切替方式

二重化したNIC(LANカード)を同一ネットワーク上に接続し、排他使用して伝送路の切替えを制御します。通信相手が限定されず、またルータを経由した別ネットワーク上のホストとの通信も可能です。

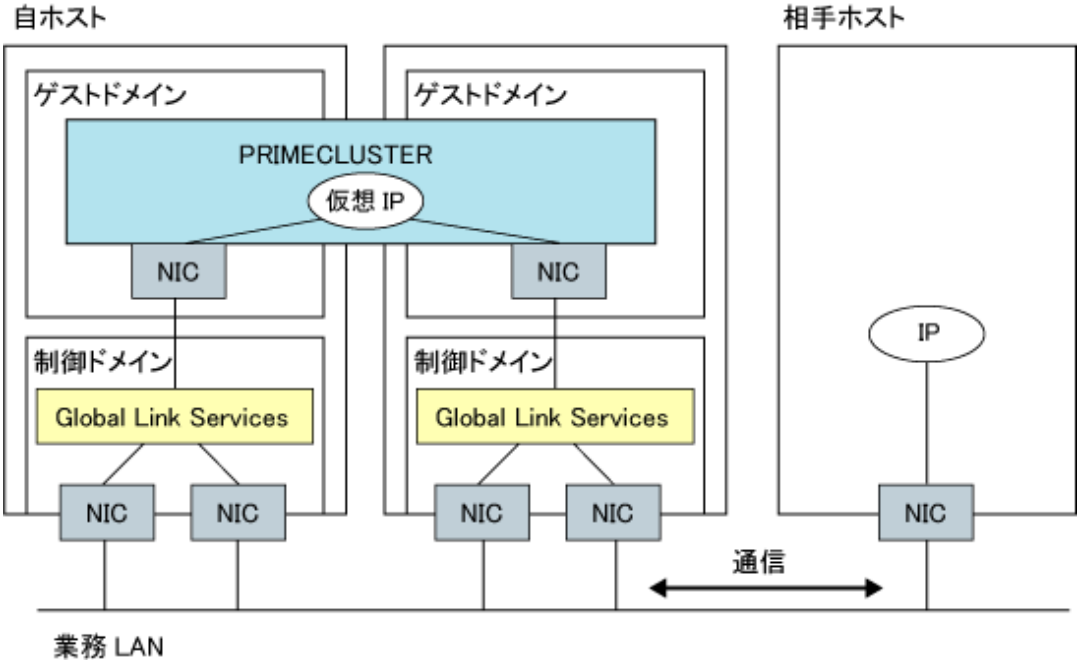
図2.14 NIC切替方式



2.3.10.3 仮想NIC方式（Solaris）

同一ネットワーク上に接続した複数の物理NIC(LANカード)を、論理的に1本に見せるための仮想的なインタフェースを生成して通信を行います。通信相手が限定されず、またルータを経由した別ネットワーク上のホストとの通信も可能です。Oracle VM環境の場合、制御ドメイン上で作成した仮想インタフェースをゲストドメインの通信に使用することが可能です。

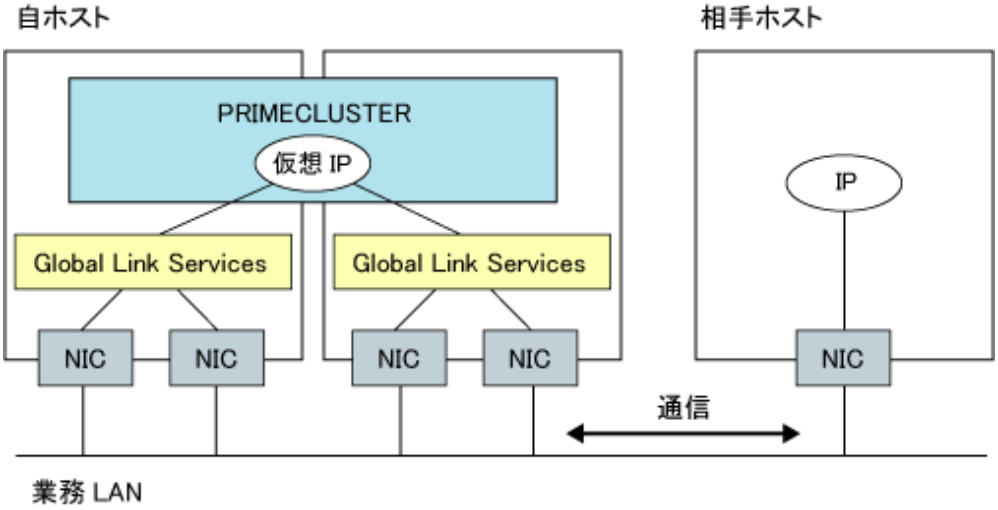
図2.15 仮想NIC方式（Solaris)



2.3.10.4 仮想NIC方式（Linux）

同一ネットワーク上に接続した複数の物理NIC(LANカード)を、論理的に1本に見せるための仮想的なインタフェースを生成して通信を行います。本方式では、二重化したNICを排他使用して伝送路の切替えを制御します。通信相手が限定されず、またルータを経由した別ネットワーク上のホストとの通信も可能です。

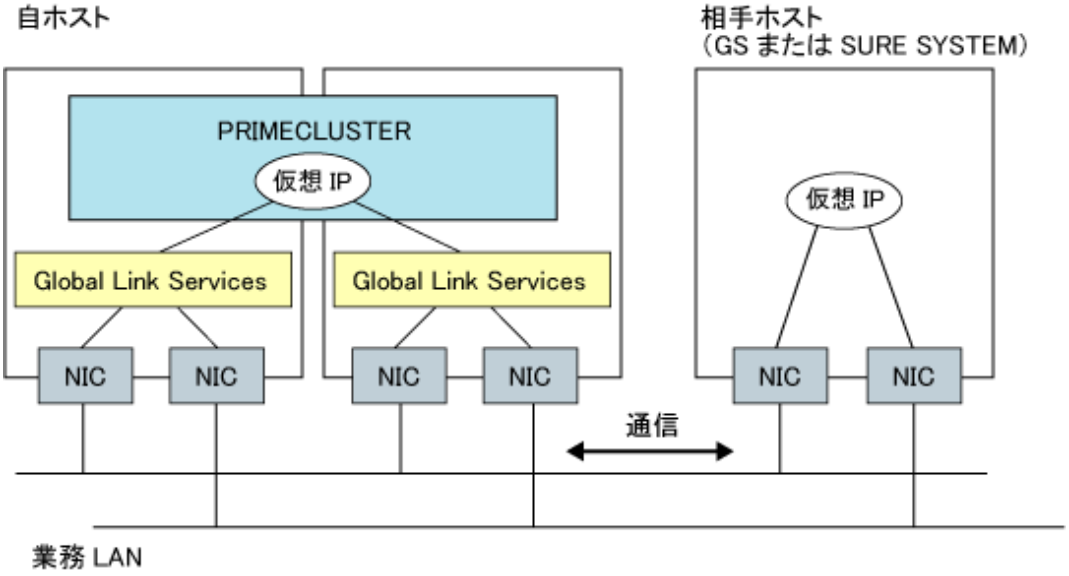
図2.16 仮想NIC方式（Linux)



2.3.10.5 GS/SURE連携方式（Solaris）、GS連携方式(Linux)

グローバルサーバとの間で高信頼通信を行うための富士通方式に従って伝送路を制御します。本方式では二重化した伝送路を同時に使用し、正常時はTCPコネクションごとに伝送路を自動的に振り分けて通信を行い、異常発生時には、該当の伝送路を切り離してTCPコネクションを正常な伝送路へ移動し、縮退運用を行います。

図2.17 GS/SURE連携方式（Solaris）、GS連携方式(Linux)



第3章 クラスタインタコネクトの詳細

本章では、クラスタインタコネクト概要と要件について説明します。

3.1 概要

クラスタインタコネクトとは、PRIMECLUSTERクラスタシステムの最も基本的な構成要素です。PRIMECLUSTERが提供する全てのサービスは、各ノード間の通信を行い、ハートビート要求の応答によりノード状態を判断するためのクラスタインタコネクトに依存しています。

3.1.1 クラスタインタコネクトと通常のネットワークとの違い

クラスタインタコネクトは従来のネットワークとは用途が異なります。クラスタインタコネクトの最も重要な役割は、ハートビート要求の送信および応答です。

ハートビートメッセージは、ノードおよびクラスタインタコネクトの状態を判断するために使用されます。メッセージが到達しない場合、クラスタソフトウェアは障害が発生したと判断してリカバリ処理を実行します。ただし、100パーセント信頼できるネットワークハードウェアは存在しないため、PRIMECLUSTER ICFプロトコルは、パケットの紛失や配送順序エラーなどのエラーを、ある範囲までは容認します。

クラスタインタコネクトでは、物理的に接続されているネットワークを冗長化することにより、1箇所で障害が発生してもメッセージを送信できるようにしなければなりません。ただし、全てのクラスタインタコネクトが故障した場合は、クラスタ管理ソフトウェアはノード障害が発生したときと同様にリカバリ処理を実行します。

3.1.2 インタコネクトプロトコル

PRIMECLUSTERが使用するICF (ノード間通信機構) プロトコルは、クラスタ通信専用に設計されています。この高速プロトコルはメッセージの順序正しい配送を保証します。ICFはTCP/IPよりオーバーヘッドが少なくなります。ICFはイーサネットプロトコルまたはサービスオーバーIP (CF/IP) を使用します。

注意

- TCP/IPのみをサポートしているデバイスでは、直接イーサネット上に設定されたICFプロトコルをルーティングすることはできません。この場合、ルーティングにはレベル2のルータを使用する必要があります。
- ICFはCFの内部コンポーネントのみで使用可能であり、ユーザレベルリソースで使用することはできません。クラスタインタコネクトにアクセスするアプリケーションにはクラスタインタコネクトプロトコル (CIP) を使用します。CIPはICF上で標準的なTCP/IPプロトコルを提供します。

3.2 クラスタインタコネクトの要件

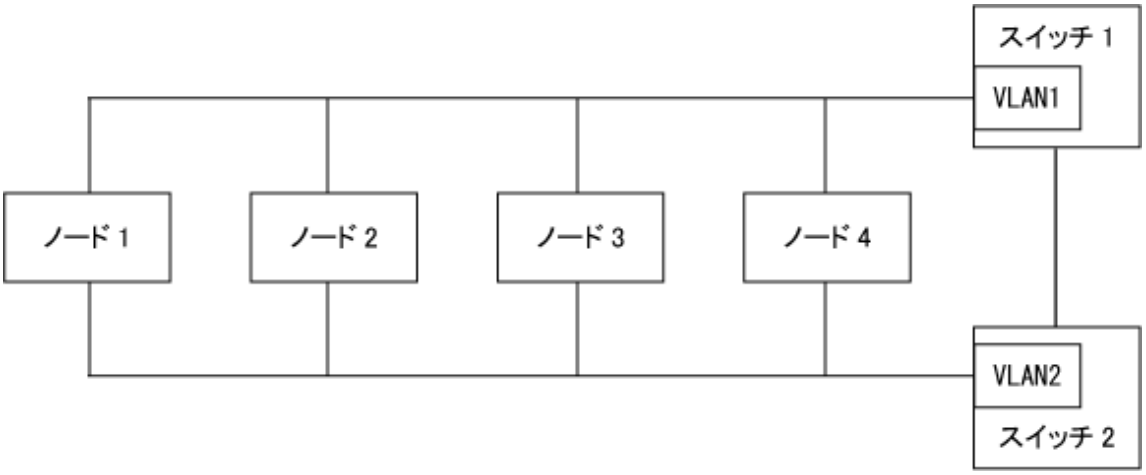
PRIMECLUSTERは、イーサネットデバイスおよび、TCP/IPをサポートするデバイスをクラスタインタコネクトとして使用します。クラスタインタコネクトは冗長化して、クラスタ内の全てのノード間に複数の独立した接続を装備する必要があります。

クラスタインタコネクトで使用する複数のスイッチ間、ルータ間などを接続する場合は、VLANを使用し、それぞれの単位(スイッチ単位など)で論理的に切り離してください。

例

以下の図のように複数のスイッチの間を接続する場合は、VLAN を使用し、スイッチ単位で論理的に切り離します。

図3.1 複数のスイッチの間を接続する場合



また、各クラスタインタコネクトは、クラスタシステムの全ノードに接続されていなければなりません。

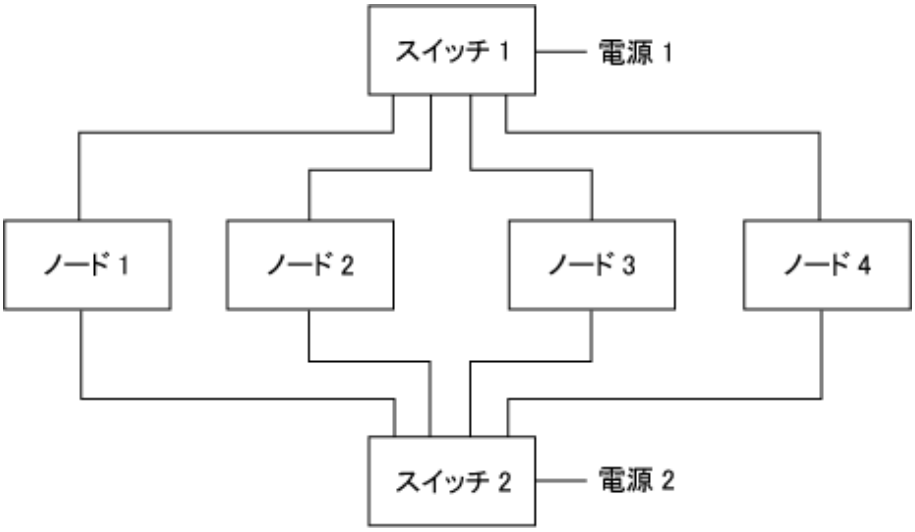
3.2.1 冗長化

クラスタインタコネクトを冗長化するには、複数の独立した接続と、独立した経路を用意する必要があります。冗長クラスタインタコネクトの例を以下に示します。

- 各イーサネットボードで、使用するポートは1つのみにする。1つのボードに複数のポートが存在するボード(Quad Fast Ethernet Cardなど)を使用すると、ボードの一点故障がクラスタインタコネクトの全系停止を引き起こすことになるため、各ノードへの接続を独立させる。
- 複数のクラスタインタコネクトに同じハブ、スイッチ、ルータを使用しない。

以下の図は、一般的な4 ノードのクラスタシステムを示しています。この図には2つのスイッチがあり、各スイッチの複数回線はそれぞれ固有の電源に接続されています(2つのスイッチの間を接続する場合は、VLANを使用し、スイッチ単位で論理的に切り離す必要があります)。スイッチがラック内で同じ電源装置に接続されていると、電源装置の一点故障時に、クラスタインタコネクトの全系停止が発生する可能性があります。

図3.2 一般的な4ノードクラスタシステム



3.2.2 経路

クラスタシステムが信頼性を保証するには、クラスタインタコネクトの冗長化が不可欠です。このため、ICFは使用可能な帯域幅を全て使用するように設計されています。クラスタシステムのノード間の各々の接続を経路と呼び、ICFが他のノードメッセージを送信する時、使用可能な全ての経路間でトラフィックを分散するように、経路を選択します。図3.2 一般的な4ノードクラスタシステムに示す4ノードクラスタの場合、各ノードには他の各ノードにつながる2つの経路があります。

3.2.2.1 ハートビート

クラスタインタコネクトの動作が全て正常であれば、経路は有効 (UP状態) となります。冗長化されたクラスタインタコネクトが同じ速度で動作しているものであれば、全ての経路がメッセージ通信に使用され、UP状態となります (速度の異なるクラスタインタコネクト (100Mイーサネットとギガバイトイーサネットの混在など) については後述します)。さらに、ハートビート要求には常に冗長化されたクラスタインタコネクトの全経路が使用されます (ハートビートとは、ノードが機能していることを示すメッセージです)。ある経路上でハートビート要求が失敗した場合、その経路はDOWN状態と判定されます。DOWN状態の経路はPRIMECLUSTERの系間通信に使用されません。ただし、DOWN状態の経路でハートビート要求がリトライされる場合がありますが、この要求が成功すれば、DOWN状態の経路はUP状態に自動復旧します。このように、系間通信にUP状態の全経路を使用することをポートアグリゲーション (Port Aggregation) またはトランキング (Trunking) といいます。

ハートビートが失敗する原因は複数考えられます。ネットワークコンポーネントに障害が発生すると、その経路は使用できなくなり、上記の動作が起こります。特定ノードと他のノードとの最後の経路がDOWN状態になったときは、その経路にDOWNマークが付くのではなく、ノードがLEFTCLUSTERとなります。

LEFTCLUSTERとは、ノードが同じクラスタにある他のノードと通信できないことを示す状態をいいます。これは、ノードがクラスタに参入していないことを意味します。ノードがLEFTCLUSTERまたはDOWN状態になると、クラスタに再参入するまで、そのノードに対するハートビートは行われなくなります。

注意

ノードがLEFTCLUSTERまたはDOWN状態であっても、最後の経路はDOWN状態にはなりません。

ノードの障害でなくてもノードがハートビート要求に応答することを妨げるイベントが発生する場合があります。たとえば、リンク修復の試行中にファイバチャネルコントローラが他のネットワークデバイスドライバの実行を妨げる場合があり、これがノード障害と誤認される可能性があります。このような場合に備えて、最後の経路上でハートビート要求が何回失敗したらノードをLEFTCLUSTERにするかを示すパラメタはチューニング可能です。

ICFのその他のパラメタをチューニングすることはできません。経路検出アルゴリズムでは、できるだけ早く経路をDOWN状態として、メッセージを他の経路に切替えて、著しい遅延をもたらさないようにすることが重要であるからです。ある経路が間欠故障などの理由で一時的に使用できない状態でも、自己復旧方式により経路が動作可能と判断されれば、その経路が使用可能 (UP状態) になります。ICFパラメタの調整については、“PRIMECLUSTER Cluster Foundation 導入運用手引書 ” を参照してください。

PRIMECLUSTERは、ギガビットイーサネットと100Mイーサネットとの組み合わせのような、速度の異なるクラスタインタコネクトの組み合わせもサポートしています。上記の例の場合、PRIMECLUSTERは系間通信に100Mイーサネットよりは、ギガビットイーサネットを使用するように経路を選択することになります。ポートアグリゲーション (Port Aggregation) /トランキング (Trunking) は同速度の全てのデバイスに対して使用可能であり、非対称の組み合わせの場合は使用されません。しかし、ハートビート要求はクラスタインタコネクトの速度に関係なく、常に全てのクラスタインタコネクトに送信されます。

3.2.3 設計時の検討項目

クラスタインタコネクトをどのように冗長化して使用するか、どのデバイスを使用するかといった設計では、以下の事項を検討する必要があります。

- 帯域幅
- 応答待ち時間 (レイテンシ)
- 信頼性
- デバイスインタフェース
- セキュリティ

3.2.3.1 帯域幅

PRIMECLUSTER自体は多くの帯域幅を必要とはしていません。PRIMECLUSTERの各クラスタインタコネクトに必要な帯域幅は0.002Mbps未満です。
このため、以下の場合は、帯域に関する考慮は必要ありません。

- 業務LANや管理LANとクラスタインタコネクトが1つの帯域を共有しない場合
- ユーザアプリケーションがクラスタインタコネクトを使用して通信を行わない場合

帯域幅の使用例については、以下の表を参照してください。“図3.2 一般的な4ノードクラスタシステム”の構成では、クラスタインタコネクトに100Mbpsイーサネットが2つ構成されています。各クラスタインタコネクトが使用できる帯域幅は80 Mbps、各ノード上のエンドユーザアプリケーションがクラスタファイルシステムおよびその他のアクティビティに使用する帯域幅は36Mbpsであるとします(これはあくまで使用例です。実際に使用する帯域幅はアプリケーションによって異なります)。

表3.1 2つの100Mbpsイーサネットボードによるクラスタインタコネクトの例

項目	帯域幅	合計帯域幅
100Mbps イーサネット×2	80Mbps	160Mbps (=2インタコネクト×80Mbps)
PRIMECLUSTER要件	0.002Mbps	0.016Mbps (=4ノード×2インタコネクト×0.002Mbps)
ユーザ業務要件	36Mbps	144Mbps (=36Mbps×4ノード)

```
合計使用率 = (PRIMECLUSTER要件 + ユーザ業務要件) / 100Mbpsイーサネットの合計帯域幅 × 100
          = (0.016 + 144) / 160 × 100 = 90%
```

この例では、2つの高速イーサネットインタコネクトが帯域幅の90パーセント以上を使用していることになります。

注意

合計使用率が100パーセントに近いとクラスタインタコネクトの応答待ち時間が増加し、ハートビート切れを誤検出する可能性があるため、初期設定時には、30パーセント以上の余裕を残しておくような設計をすることを推奨します。

この例の構成と、それにかかる負荷状況から、1つの高速イーサネットインタコネクトを増設して余分な容量を確保することが推奨されます。以下の表は増設後の計算結果を示しています。

表3.2 3つの100Mbpsイーサネットボードによるクラスタインタコネクトの例

項目	帯域幅	合計帯域幅
100Mbps イーサネット×3	80Mbps	240Mbps (=3インタコネクト×80Mbps)
PRIMECLUSTER要件	0.002Mbps	0.024Mbps (=4ノード×3インタコネクト×0.002Mbps)
ユーザ業務要件	36Mbps	144Mbps (=36Mbps×4ノード)

```
合計使用率 = (PRIMECLUSTER要件 + ユーザ業務要件) / 100Mbpsイーサネットの合計帯域幅 × 100
          = (0.024 + 144) / 240 × 100 = 60%
```

増設後の新しい構成では、帯域幅に40パーセントの余裕があり、30パーセント以上の余裕を確保するべきという推奨例を満たすことになります。本例では3重に冗長化されたクラスタインタコネクトを使用していますが、PRIMECLUSTERは最大4重のクラスタインタコネクトをサポートしています。

3.2.3.2 応答待ち時間 (レイテンシ)

前述してきたように、PRIMECLUSTERはハートビートの要求および応答により、ノードおよびその他のリソースが正常に動作しているかどうかを判断します。一定時間内にハートビートの応答がなければ、PRIMECLUSTERはリカバリ処理を開始します。各ノードのCluster Foundation (CF) は、各クラスタインタコネクト上で、クラスタシステムを構成する自分以外の全てのノードに、200 msごとにハートビートを送信します。タイムアウト(デフォルト10秒)までに、ハートビートの要求を200 msに一度×50回試行しても、相手ノードからすべて応答がなければ、CFはそのノードをLEFTCLUSTER状態になったと判断します。

200 msはクラスタインタコネクトの応答待ち時間の上限として、サイズの小さいメッセージや応答を長距離伝送するのには十分な間隔として設計されています。この間隔は固定値であり変更することはできません。

3.2.3.3 信頼性

イーサネットをPRIMECLUSTERのクラスタインタコネクトとして使用する場合は何の問題もありません。PRIMECLUSTERの通信プロトコルはICFであり、ICFは送信先に正確かつ順序正しくメッセージ送信を行うことを保証します。ただし、ICFは信頼性の高い通信を重点において動作します。そのため、クラスタインタコネクトの信頼性が高い場合は、ICFのオーバーヘッドはほとんどありませんが、クラスタインタコネクトの信頼性が低い場合、ICFのオーバーヘッドが増加します。TCP/IPなどのプロトコルの場合と同様、クラスタインタコネクトにエラーが発生すると、メッセージの再送が行われます。

注意

- メッセージの再送には帯域幅が使用され、かつ、応答待ち時間にも影響するため、再送が発生しないように信頼性の高いクラスタインタコネクトを使用することが重要になります。
- イーサネットのエラー率が1/1,000,000バイトより多い場合、イーサネット層にて調査する必要があります (エラー率を調べるには、netstat(1)またはip(1) コマンドを使用します)。

3.2.3.4 デバイスインタフェース (Solaris)

PRIMECLUSTERはDLPI (Data Link Provider Interface) を使用します。デバイスがDLPIをサポートしていない場合は、PRIMECLUSTERはそのデバイスをクラスタインタコネクト対応のデバイスとして認識することができません。さらに、クラスタインタコネクト対応デバイスとして認識されるためには、イーサネットデバイスとしてOSに認識されている必要があります。TCP/IPをサポートしていてもイーサネットでないデバイスもありますが、PRIMECLUSTERのクラスタインタコネクトに使用されるプロトコルはTCP/IPではなくイーサネットであることを認識しておいてください。

3.2.3.5 セキュリティ

PRIMECLUSTER製品は、クラスタインタコネクトを専用のネットワークにすることを想定していますが、ICFは物理媒体上で動作する他のプロトコルと干渉しないため、業務LANを使用することは技術的には不可能ではありません。しかし、PRIMECLUSTERのセキュリティモデルはクラスタインタコネクトを構成するネットワークを、物理的に業務LAN他から切り離すことによって実現します。

注意

セキュリティ上の理由も含め、クラスタインタコネクトに業務LANを使用しないでください。

クラスタインタコネクトに業務LANを使用すると、PRIMECLUSTER製品がインストールされていれば、業務LAN上のどのマシンでもクラスタに参入することができてしまいます。これにより不正なユーザが参入してクラスタサービスにフルアクセスすることも可能となってしまうからです。

第4章 RMS (Reliant Monitor Services)

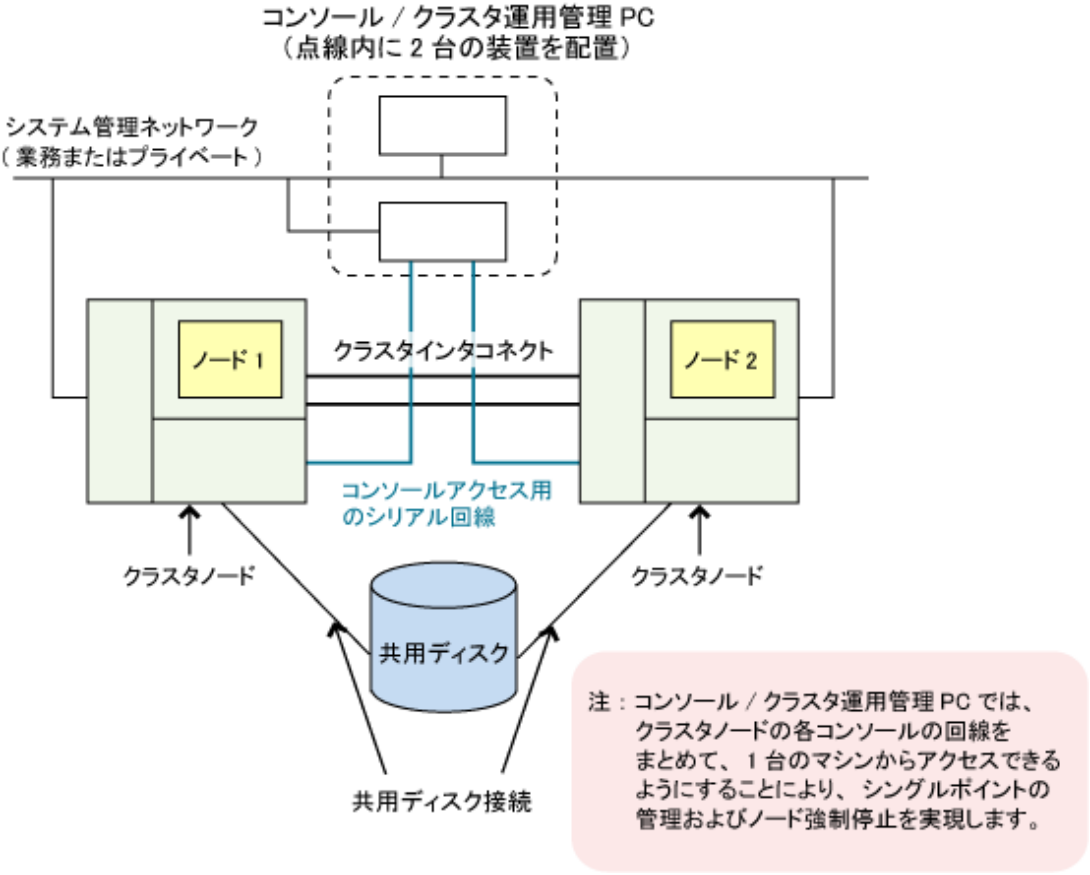
本章では、RMSの基本的な概念について説明します。まず、RMSの高可用性を実現するために用いられる基本的な用語と概念から始めて、その後にRMSの動作の詳細について説明します。

4.1 RMSの概要

RMSは、ユーザ業務に対して高可用性を実現するように設計されたソフトウェアです。最小構成として、共用ディスク装置に接続されている2ノードクラスタシステムは、コンソールまたはクラスタ運用管理PCから管理されます。RMSは、ディテクタによりコンポーネントの状態や、ユーザ業務に必要とされるリソースを監視します。ユーザ業務や、ユーザ業務を構成するリソースの障害が発生すると、その障害をディテクタが検出し、RMSへ通知します。通知を受けたRMSは、次のセクションで説明するような処置を行います。

以下の図では、Solaris上の基本的なRMSクラスタコンポーネントについて説明します。

図4.1 Solaris上のRMSクラスタ



RMSの高可用性とは、各ノードを停止させないということではなく、ユーザ業務の可用性を最大限に引き出すことを意味しています。

RMSで高可用性を実現する方法には、冗長化と切替えの2種類があります。

4.1.1 冗長化

高可用性を保証するために、以下の冗長手法が使用されます。

- 複数台のノードでそれぞれが他のクラスタノードのリソース負荷を引き継げるように構成
- ディスクのミラー化、ハードウェアRAID、リモートミラー化によるデータの複製
- 記憶装置へのアクセスのパスの冗長化
- ノード間通信専用の通信経路の冗長化

RMSによる冗長化システムは以下の要素により構成されます。

複数ノード

同一のオペレーティングシステムおよびRMSソフトウェアを導入した複数ノードによるクラスタを構成します。RMS構成でサポートする最大ノード数は、理論上無制限です。

参照

適切なノード数については、“PRIMECLUSTER ソフトウェア説明書 / インストールガイド”を参照してください。

共用記憶装置

RMS構成で定義されている全ノードで共用している、共用ディスク装置上の全てのデータにアクセスできるようにしておく必要があります。このためには、通常、全てのノードが SANの共用ディスクにアクセスできるようにしておく必要があります。ただし、NASなどの他のアクセス方法を使用することも可能です。

RMSネットワーク

RMSでは、RMS構成で定義されているノード間の通信に、TCP/IPプロトコルを使用します。RMSは、RMS構成で定義されている他ノードのRMSを監視するためにPRIMECLUSTERを構成する冗長化されたクラスタインタコネクト上のCluster Interconnect Protocol (CIP) を使用します。

参照

詳細については、“3.1.2 インタコネクトプロトコル” を参照してください。

4.1.2 アプリケーションの切替え

RMSはオブジェクト指向に基づいて設計されています。仮想ディスク、ファイルシステムのマウントポイント、プロセスなどのあらゆるシステムコンポーネントが「オブジェクト」として認識されます。これらの「オブジェクト」は「RMSリソース」と呼ばれます。「RMSリソース」はオブジェクトタイプ (Object Type) と呼ばれる単位グループに分類され、オブジェクトタイプにはプロパティと呼ばれる属性 (Attribute) があり、この属性によりRMSリソースに対して実行/監視または異常発生時のリカバリ処理を限定または定義します。RMSリソースはディテクタと呼ばれるプログラムにより監視されます。このようにきわめて汎用化されたオブジェクト指向設計により、監視の種類やレベルに高度な柔軟性をもたらすことができます。

また、互いに依存関係にあるリソースをグループ化して論理的なグループを設定することができます。グループ内のRMSリソースに障害が発生すると、ユーザ業務全体が障害に対応することができます。

4.1.2.1 自動切替え

RMSリソースの障害が検出されると、ユーザが定義した処理が実行されます。多くの場合、障害に対する最大の対処法は切替処理です。

切替処理はフェイルオーバとも呼ばれ、最初にユーザ業務をOffline状態にしてから、RMSの制御により他のノード上でユーザ業務を再起動します。RMSでは、対称的な切替えの定義が可能です。たとえば、全てのRMSノードが他の任意のRMSノードからリソースを引継ぐことができるように設定することが可能です。これにより、ユーザ業務を実行中のノードに障害が発生すると、RMSは自動的にそのノードを停止して、ユーザ業務を稼動中の他のノードに切替えます。

自動切替えの詳細は、障害の発生時にRMSが起動するユーザ定義スクリプトおよび構成定義ファイル (usファイル) に定義されます。このファイルの作成には、RMSウィザードが使われます。

4.1.2.2 手動切替え

RMSで監視しているユーザ業務を手動切替えすることで、ハードウェアのメンテナンスなどを行うことができます。たとえば、2ノードのRMS構成では、1台のノードのメンテナンスを行う場合は、全てのユーザ業務を一時的にもう1台のノードに切替え、ノードのメンテナンスを行います。メンテナンス完了後、もう1台のノードに対するメンテナンスを行うために、ユーザ業務の切戻しを行います。なお、RMSの切替えを実行している間、ユーザ業務が提供するサービスが瞬間的に中断する可能性があります。また、全てのアプリケーションが1台のノード上で動作している間は、負荷が集中することにより、応答時間が低下する可能性があります。

4.1.2.3 IPエイリアス

RMSではIPエイリアスを使うことにより、IPアドレス切替えを可能としています。これにより、1つの物理ネットワークインタフェースに対して、複数のIPアドレス(IPエイリアス)を割当てることができます。このIPエイリアスを行うことで、RMSの切替えの結果、他のノードでユーザ業務を実行する場合にも、ユーザは同じIPアドレスで通信を続けることができます。この機構により、PRIMECLUSTERはIPアドレス引継ぎを可能とします。

4.1.2.4 データの整合性

RMSにより、ユーザ業務の切替えを実行する場合には、RMSは現在のノード上のユーザ業務に関連する全てのRMSリソースをOfflineにしてから、次にこれらのRMSリソースを新しく稼動させるノード上でOnlineにします。この手順を踏むことにより、複数ノードが同じ共有資源に対するアクセスの競合により、データが破損する危険を回避することができます。

複数の障害が同時に発生(二重故障または三重故障)した場合には、RMSは切替え自体を禁止して競合によるデータの破損を防ぎます。そのため、特定の状況では切替えが全くできなくなるケースもありえます。

RMSでは高可用性を目標にしていますが、データが破損する恐れがある場合は高可用性よりもデータの整合性を優先します。

4.2 RMSの監視と切替え

RMSは、RMSリソースの高可用性を実現するために、以下のプロセス、スクリプト、ファイル、パラメタで構成されます。

- BM (ベースモニタ)
- 構成定義ファイル
- 構成スクリプト
- ディテクタ
- RMS環境変数

構成スクリプト、ファイル、および環境変数はカスタマイズ可能であり、切替えに対する顧客固有のニーズに合わせて調整することができます。

参照

カスタマイズの詳細については、“4.4 カスタマイズオプション”を参照してください。

4.2.1 BM (ベースモニタ)

BM (ベースモニタ) は、RMSクラスタからノードを監視するための中心的なプロセスです。BMは以下の機能を実行します。

- RMSの全ての状態変更を制御および調整
- 監視中のリソースに問題が発生した場合に定義に従いリカバリ処置の実施
- RMSリソースに関する情報を取得し、システム管理者がRMSの使用要件に応じて作成したRMS 構成定義ファイルからリカバリ処理を実行

BMはCluster Admin GUIまたはhvcmコマンドで起動され、クラスタの各ノード上でbmという名前のプロセスとして実行されます。

4.2.2 構成定義ファイル

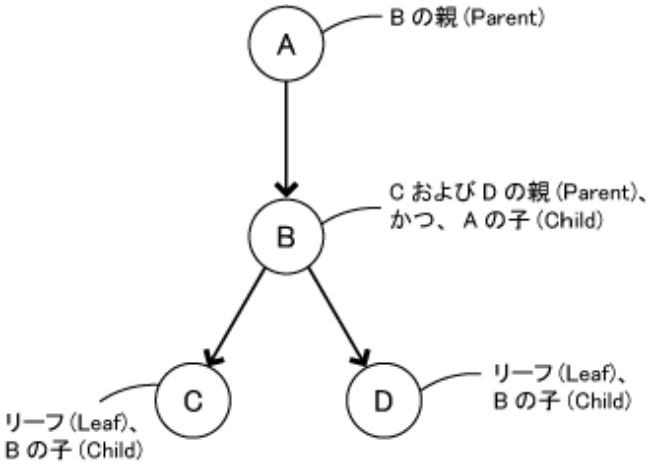
RMS構成定義ファイルは通常、RMS Wizard Toolsによって生成されるテキストファイルで、以下に記載するRMSで監視するリソースオブジェクトの定義やオブジェクトの相互依存性などのRMSオブジェクトの構成が格納されます。RMS構成定義ファイルの記述内容はRMS実行時にBMに渡されます。

4.2.2.1 オブジェクト間の依存関係

RMSでは、親 (Parent) および子 (Child) という用語は、オブジェクトおよびRMSリソースオブジェクト間の依存関係を示します。リーフオブジェクト (Leaf Object) という用語は、子が存在しないシステムグラフのオブジェクトを示します。リーフオブジェクトは構成定義ファイルの最初に定義されます。リーフオブジェクトは子を持つことができませんが、子は親および子を持つことができます。

以下の図は、オブジェクト間の親子関係を示しています。

図4.2 オブジェクトの親子関係



上図は、以下の関係を示しています。

- オブジェクトAが正常に動作するためには オブジェクトBが必要。オブジェクトAはオブジェクトBの親 (Parent)であり、オブジェクトBはオブジェクトAの子 (Child) である。
- オブジェクトBが正常に動作するためにはオブジェクトCおよびオブジェクトDが必要。オブジェクトBはオブジェクトCおよびオブジェクトDの親 (Parent) である。
- 子 (Child) が存在しないオブジェクトCおよびオブジェクトDはリーフオブジェクト (Leaf Object) である。オブジェクトCおよびオブジェクトDは、オブジェクトBの子でもある。

4.2.2.2 オブジェクトタイプ (Object Type)

オブジェクトタイプ (Object Type) とは、監視方法等を共有する同種のリソースグループとしてまとめたものです (例として、グループ内の全てのオブジェクトが同じディスク装置を使用する場合などがあります)。グループ内の全てのオブジェクトに対する各オブジェクトタイプには、属性 (Attribute) が存在し、この属性により RMS リソースに対して実行する監視の種類を限定または定義します。つまり、特定のオブジェクトタイプに関連付けられた属性を使って、RMS リソースに対するBMの動作方法を定義することができます。属性は一般にデバイス名やスクリプトを指定するために使用され、RMSリソースのオブジェクト定義 (Object Definition) により任意の順序で指定することができます。

属性には必須属性と任意属性があります。必須属性の例として、vdiskオブジェクトタイプのデバイス名の指定があります。また、一般的な任意属性の例として、ある特殊な条件下で実行されるようなスクリプトがあります。多くの属性は、ほとんどのオブジェクト定義に使用されます。

一方、特定のオブジェクトタイプのみに有効な属性や、オブジェクト定義に特定の属性を使用することが必須とされるオブジェクトタイプもあります。

4.2.2.3 オブジェクト定義 (Object Definition)

オブジェクト定義 (Object Definition) とは構成定義ファイルに記述されるステートメントで、“object” というキーワードで始まります。このキーワードにより、RMSが認識できる用語で特定のリソースを指定します。オブジェクト定義に指定する項目を以下に示します。

- オブジェクトタイプ (Object Type)
- RMSリソース名
- 属性 (Attribute)
- 特定のRMSリソースが依存している子 (Child) RMSリソース

RMSのインストールおよび確認が完了したら、RMSリソースの監視を開始する前に、構成定義ファイルを設定する必要があります。構成定義ファイルのオブジェクト定義に指定されていない対象に対しては、RMSリソースとしてRMSに認識されることはありません。

4.2.3 構成スクリプト

RMS構成スクリプトは、RMSリソースの状態の変化に反応したり、状態の変化を呼び出したりするシェルプログラムまたは実行可能ファイルの集まりです。

RMSリソースの状態には以下の種類があります。

- Faulted
- OfflineFault
- Offline
- Online
- Unknown
- Wait
- Deact
- Inconsistent
- Standby
- Warning

RMSの活性処理またはリカバリ処理は、全てのスクリプトにより実行されます。スクリプトを使用しない場合、RMSリソースの監視は可能ですが、どんな状態変化による活性処理も行われません。スクリプトは特定リソースのオブジェクト定義 (Object Definition) の属性 (Attribute) として認識され、また、ディテクタから状態の変化に応じてBMから実行されます。たとえば、RMSネットワークリソースの状態がOnlineからFaultedに変わると、BMはRMSリソースのオブジェクト定義に定義されているFaultScriptを起動します。

RMSのスクリプトは、状態を変更するスクリプト (要求トリガ・スクリプト) と状態に反応して動作するスクリプト (状態トリガ・スクリプト) に分類されます。

要求トリガ・スクリプトを以下に示します。

- PreOnlineScript
- PreOfflineScript
- PreCheckScript
- OnlineScript
- OfflineScript
- OfflineDoneScript

状態トリガ・スクリプトを以下に示します。

- PostOnlineScript
- PostOfflineScript
- FaultScript

使用しているRMSによっては、追加スクリプトが含まれている場合もあります。

参照

RMSリソース状態およびスクリプトの詳細についてはマニュアル、“PRIMECLUSTER RMS 導入運用手引書” を参照してください。

4.2.4 ディテクタ

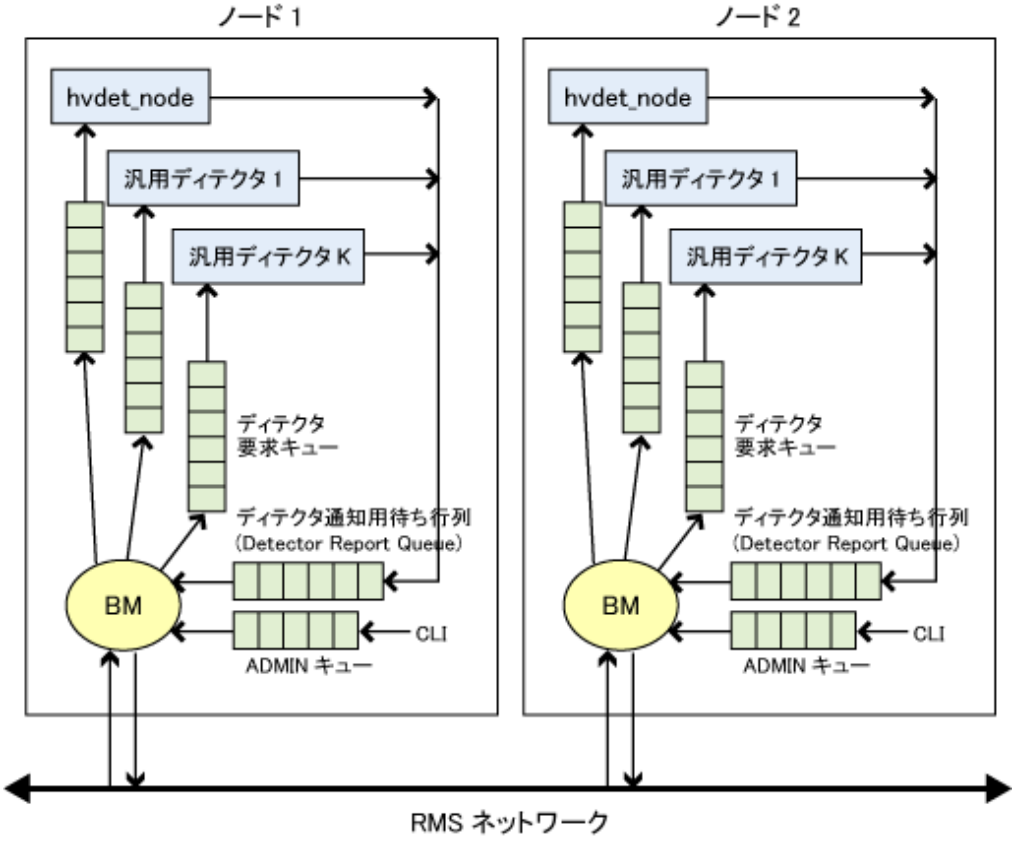
ディテクタとはリソースの状態を監視してベースモニタにリソースの状態を通知するものです。

各ディテクタは1つ以上のリソースに関する情報を内部テーブルで管理します。ディテクタは内部テーブルに記録されている各RMSリソースを検索し、各RMSリソースの最新の状態を取得して、最新の状態と前回の状態を比較します。最新の状態と現在の状態が異なる場合、

ディテクタは該当するRMSリソースタイプのディテクタ通知用待ち行列 (Detector Report Queue) により、BMに最新の状態を通知します。BMは、構成定義ファイルからの情報に基づいてリソースの状態に応じた処理を実行します。

以下の図は、BMとRMSディテクタ間の通信フローを示しています。

図4.3 BM (ベースモニタ) の通信フロー



4.2.5 RMS環境変数

ユーザはRMS環境変数を使って、BMに起動時にBMを対して値を指定します。RMSには可用性を高めるためのHV_AUTOSTART_WAIT、RELIANT_PATHなどの数多くのRMS環境変数があります。RMS環境変数のデフォルト設定は各クラスタシステムおよびユーザ業務に応じて調整することができます。

4.3 RMSの管理

RMSの管理には2種類の方法があります。1つはコマンドラインインタフェース (CLI) によるもので、もう1つはグラフィカルユーザインタフェース (GUI) によるものです。Cluster Admin GUIを使用することを推奨します。

4.4 カスタマイズオプション

障害の検出およびリカバリ方法を、ユーザ要件、システム環境に応じてカスタマイズすることができます。このカスタマイズは、RMSの監視対象リソースの構成定義ファイル、ディテクタ、スクリプトを変更することにより行います。構成変更は、あらかじめ事前に計画/設計をしておいて、十分なテストを実施した上で、カスタマイズする必要があります。

4.4.1 汎用リソースタイプとディテクタ

RMSには、IPエイリアス、ファイルシステムなどの、通常のシステムレベルのリソースタイプでは使用できないような専用リソースを定義するための、汎用リソースタイプがあります。このような汎用リソースを監視するための、汎用ディテクタは、最大64種類のリソースタイプおよびディテクタを定義することができます。

第5章 RMSウィザード

本章では、RMSウィザードの基本的な概念について説明します。まず概要を説明し、次にRMSウィザード製品とその機能について説明します。

5.1 RMSウィザードの概要

複数の顧客アプリケーションを組み合わせて、顧客環境に合わせたRMS構成を作成するのは、非常に複雑な作業であるため、ユーザ業務で使用する各アプリケーションと、RMS環境に詳しい専門家が必要になります。しかし、RMSウィザードを使えば、RMSの構成設定および運用が簡単に行うことができます。

注意

RMSウィザードはシステム、RMS、およびアプリケーションを管理します。ただし、RMS、RMS Wizard Toolsは別コンポーネントの扱いをとります。

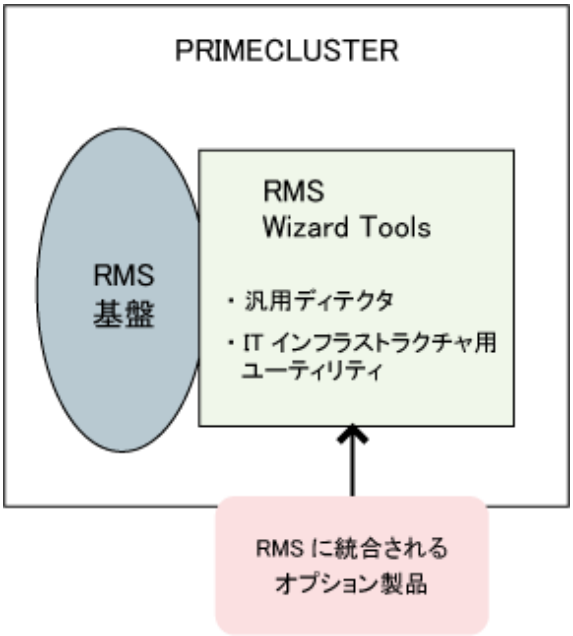
5.2 RMSウィザードのアーキテクチャ

RMSウィザードは以下の製品で提供されます。

- RMS Wizard Tools
  RMSウィザードのフレームワークおよびRMS BM (ベースモニタ) とのインタフェースを提供します。RMS構成の設定が簡素化され、RMSとの統合によりHAクラスタ環境が強化されます。

以下の図は、RMS Wizard ToolsとRMSとの関連性を示しています。

図5.1 RMSウィザードのアーキテクチャ



5.3 RMS Wizard Tools

オプション製品であるRMS Wizard Toolsは、高可用性のRMS構成を作成および管理する製品です。RMS Wizard Toolsには以下の機能があります。

- クラスタ内のノード上で構成設定済みのユーザ業務の起動
- 構成全体をコピー

・ 構成の変更を許可または禁止

さらに可用性の高い構成にするには、RMS Wizard Toolsのユーティリティウィザードを使用します。ユーティリティウィザードは個別のアプリケーションウィザードではサポートされていない、以下のようなシステムの標準的なリソースおよびアプリケーションを管理します。

・ ファイルシステムのマウントとアンマウント

・ IPエイリアス (仮想IP割当) の設定と解除

・ 仮想ディスクの構成および構成解除

5.3.1 共用ディスク装置への対応

アプリケーションウィザードにより、RMS構成をさまざまなディスク装置用に設定できます。たとえば、GDS、Veritas VxVMなどのソフトウェアベースのソリューションを構成することも、Dell EMCなどのハードウェアソリューションを構成することもできます。このウィザードには複数のSAN (Storage Area Network) モジュールを構成する機能があります。この機能により、各モジュールを組み合わせることができます。

付録A　リリース情報

本章では、本マニュアルの主な変更箇所について説明します。

項番	版数	変更箇所	内容
1	第2版	1.2.2.1 データ整合性の保証	I/Oフェンシング機能の説明を変更しました。
2	第2版	1.7.1 Linux	仮想化対応の説明を追加しました。
3	第2版	1.7.1 Linux 2.3.5 PRIMECLUSTER SF	PRIMEQUEST の対応機種を変更しました。
4	第2版	3.2.3.1 帯域幅	帯域幅の説明を変更しました。
5	第2版	用語集	「iRMC」を追加しました。

用語集

AC (アクセスクライアント (Access Client))

アクセスクライアントを参照。

API (アプリケーションプログラムインタフェース (application program interface))

アプリケーションプログラムインタフェースを参照。

BM (ベースモニタ (base monitor)) (RMS)

RMSの中心となるリソースの可用性を管理するモジュールプロセス。BM (ベースモニタ) はデーモンとディテクタから構成され、RMSが管理するオブジェクトの状態変更の調整/制御を行う。監視中のRMSオブジェクトに異常が発生した場合には、構成定義に従ってリカバリ処理 (ローカルリカバリまたはリモートリカバリ) を実行する。

Cache Fusion

Oracle 9iで改良されたプロセス間通信インタフェース。論理ディスクブロック (バッファ) を更新する際、各ノードのローカルメモリ上にキャッシュされているブロックをディスクにフラッシュするかわりに、クラスタインタコネクト経由で、ブロックを他のノードにコピーすることで、物理I/Oのオーバーヘッドをなくし、処理を高速化することができる。

CCBR（Solarisのみ）

クラスタ構成のバックアップおよびリストアを参照。

ccbr.conf（Solarisのみ）

/opt/SMAW/ccbr ディレクトリに配置されるバックアップ/リストア用の環境設定ファイル。$CCBRHOME 変数の設定などに使用します。詳細は、cfbackup(1M)コマンドおよび cfrestore(1M)コマンドのマニュアルページおよび ccbr.conf ファイル内のコメントを参照してください。

ccbr.gen（Solarisのみ）

/opt/SMAW/ccbr ディレクトリに配置される世代数を格納するためのファイル。0 以上の値が格納されます。詳細は、cfbackup(1M)コマンドおよび cfrestore(1M)コマンドのマニュアルページを参照してください。

CCBRHOME変数（Solarisのみ）

バックアップデータが格納されるディレクトリを示します。初期値は /var/spool/pcl4.1/ccbr ディレクトリになります。この変数は、ccbr.conf ファイルでのみ設定可能です。

CF (Cluster FoundationまたはCluster Framework)

Cluster Foundationを参照。

CIM

クラスタ整合性モニタ (Cluster Integrity Monitor)

CIP

クラスタインタコネクトプロトコル (Cluster Interconnect Protocol)

CLI

コマンドラインインタフェース (command-line interface)

CLM

Cluster Manager

Cluster Foundation

基本的なクラスタリング通信サービスを提供するPRIMECLUSTERモジュールの集まり。

関連項目 クラスタ基盤

CRM

クラスタリソース管理 (Cluster Resource Management)

DLPI

Data Link Provider Interface

DOWN (CF)

ノードが使用不可であることを示すノード状態 (DOWN状態と呼ぶ)。LEFTCLUSTER状態のノードをクラスタに再参入させるためには、事前にそのノードの状態をDOWNに変更する必要がある。

関連項目 UP、LEFTCLUSTER、ノード状態

EE

Enterprise Edition

ENS (イベント通知サービス (Event Notification Services)) (CF)

イベント通知サービスを参照。

GFS 共用ファイルシステム

GFS 共用ファイルシステムは、共用ディスク装置を接続した複数ノードから一貫性/整合性を保った同時アクセスが可能であり、一部のノードがダウンしても、他のノードは処理を継続できることを特長とする共用ファイルシステムです。GFS共用ファイルシステムは、複数のノードから同時にマウントして使用できます。

Global Disk Services

ディスク装置に格納されたデータの可用性と運用管理性を向上させるためのボリューム管理機能を提供するサービス。

Global File Services

クラスタ内の2つ以上のノードから共有記憶ユニットのファイルシステムの直接、同時アクセス機能を提供するサービス。

Global Link Services

ネットワーク伝送路を冗長化することにより、ネットワークの高可用性を実現するサービス。

GUI (グラフィカルユーザインタフェース (graphical user interface))

グラフィカルユーザインタフェースを参照。

HA

高可用性 (high availability)

ICF

ノード間通信機構 (Internode Communication Facility)

I/F

インタフェース(Interface)

I/O

入出力 (input/output)

IPアドレス

インターネットプロトコルアドレスを参照。

IPエイリアス

1つの物理ネットワークインタフェースに複数のIPアドレス (エイリアス) を割り当てる機能。IPエイリアスにより、他のノードでアプリケーションを実行する場合にも同じIPアドレスで通信を続けることができる。

関連項目 インターネットプロトコルアドレス

iRMC (integrated Remote Management Controller)

PRIMEQUEST/PRIMERGYに搭載されるハードウェアの1つである integrated Remote Management Controllerの略称。

JOIN (クラスタ参入サービスモジュール (cluster join services module)) (CF)

クラスタ参入サービスを参照。

LAN (ローカルエリアネットワーク (local area network))

業務LANを参照。

LEFTCLUSTER (CF)

ノードが同じクラスタにある他のノードと通信できないことを示すノード状態。ノードがクラスタを離れていることになる。LEFTCLUSTERという中間状態は、クラスタパーティションの問題を防ぐために設けられている。

関連項目 UP、DOWN、クラスタパーティション、ノード状態

MA

非同期監視 (Monitoring Agent)

MACアドレス

MAC address。ローカルエリアネットワーク (LAN) のMAC副層で用いられる局、またはノードを示すアドレス。

MDS (メタデータサーバ (Meta Data Server))

メタデータサーバを参照。

MIB

Management Information Base

MIPC

Mesh Interprocessor Communication

MMB (Management Board)

PRIMEQUESTに搭載されるハードウェアの1つであるManagement Boardの略称。

NIC

ネットワークインタフェースカード (network interface card)

NIC切替方式

GLSが提供するLAN二重化方式の1つ。二重化したNICを排他使用し、サーバとスイッチングHUB間のLAN監視と異常検出時の切替えを実現する。

NSM

Node State Monitor

OPS (Oracleパラレルサーバ (Oracle Parallel Server))

Oracleパラレルサーバを参照。

Oracleパラレルサーバ

Oracleパラレルサーバは、クラスタ化されたプラットフォームまたはMPP (massively parallel processing) プラットフォームのユーザおよびアプリケーションにデータベースの全てのデータへのアクセス機能を提供する。

OSD (オペレーティングシステム依存 (operating system dependant)) (CF)

オペレーティングシステム依存を参照。

PAS

Parallel Application Services

PRIMECLUSTERサービス (CF)

クラスタ化アプリケーションにサービス、および内部インタフェースを提供するサービスモジュール。

PS

パラレルサーバ (Parallel Server)

RAO

RMS-Add on

RCCU (リモートコンソール接続装置 (Remote Console Connection Unit))

リモートコンソール接続装置 (Remote Console Connection Unit) の略称。

関連項目 リモートコンソール接続装置

RCI

Remote Cabinet Interface

Reliant Monitor Services (RMS)

監視、および切替機能によりユーザが指定したリソースの高可用性を維持するサービス。

RMS (Reliant Monitor Services)

Reliant Monitor Servicesを参照。

RMS Wizard Tools

RMS構成のアプリケーションの作成および管理に使用する各種設定、および管理ツールで構成されるソフトウェアパッケージ。RMSウィザードの基盤および、BM (ベースモニタ) とのインタフェースを提供する。

RMSウィザード

RMSが動作するための構成定義を作成するためのソフトウェアツール。

関連項目 RMS Wizard Tools

RMS構成

複数のノードを共用リソースに接続する構成。各ノードはオペレーティングシステム、RMSソフトウェア、固有アプリケーションのコピーを固有に保持する。

RMSコマンド

RMSリソースをコマンドラインから管理するコマンド。

SA

シャットダウンエージェント (Shutdown Agent)

SAN (Storage Area Network)

Storage Area Networkを参照。

SC

拡張性クラスタ (Scalability Cluster)

Scalable Internet Services (SIS)

Scalable Internet ServicesのTCP接続は、各接続の通常のクライアント/サーバセッションを維持しながらクラスタノード間のネットワークアクセス負荷を動的に分散する。

SCF

システム監視機構 (System Control Facility)

SD

シャットダウンデーモン (Shutdown Daemon)

SDXオブジェクト (GDS)

クラス、グループ、SDXディスク、ボリュームなど、GDSが管理する資源の総称。

SDXディスク (GDS)

GDSが管理しているディスクの総称。SDXディスクは、用途に応じてシングルディスク、キープディスク、スペアディスク、および未定義ディスクと呼ばれる場合があります。SDXディスクを単に「ディスク」と呼ぶ場合もあります。

SF

シャットダウン機構 (Shutdown Facility)

SIS (Scalable Internet Services)

Scalable Internet Servicesを参照。

Storage Area Network

複数の外部記憶装置どうしを接続し、複数のコンピュータに接続する高速ネットワーク。通常はファイバチャネルの接続。

UP (CF)

ノードが同じクラスタにある他のノードと通信できることを示すノード状態。

関連項目 DOWN、LEFTCLUSTER、ノード状態

VIP

仮想インタフェース (Virtual Interface Provider)

Web-Based Admin View

PRIMECLUSTERのグラフィックユーザインタフェースを活用するための共通基盤。インタフェースはJavaで記述されている。

WT

Wizard Tools

XSCF (eXtended System Control Facility)

eXtended System Control Facilityの略。本体装置のCPUとは独立した専用プロセッサで構成されているシステム監視機構。冷却部(FANユニット)、電源ユニット、システム状態監視、周辺装置の電源投入／切断、異常監視を一括して制御する。さらに、遠隔地からの本体装置の管理を可能にするためにシリアルポートまたはイーサネット接続経由で、本体装置をモニタする機能、故障情報をシステム管理者に通報する機能、コンソール入出力機能を兼ね備えている。

アクセスクライアント

各ノード上のGFSカーネルモジュール。メタデータサーバと通信し、共用ファイルシステムへの同時アクセス機能を提供する。

関連項目 メタデータサーバ

アプリケーションプログラムインタフェース

アプリケーションが、OSなどのサービスプロバイダが提供するサービスを利用する際に使うインタフェース。

イーサネット

IEEE802.3にて標準化されたLAN規格。現在、特殊な用途を除いて、ほとんどの LANはイーサネットである。なお、イーサネットという表現は元々10メガバイト/秒タイプのLAN規格の名称であるが、現在は高速イーサネット/ギガバイトイーサネットをも含んだ総称としても用いられる。

イベント通知サービス (CF)

クラスタ内で発生したイベントをノード間にブロードキャストする機能を提供するPRIMECLUSTERモジュール。

インストールサーバ

ネットワークを通じてクライアントマシンにオペレーティングシステムをインストールできるための設定を施したサーバ。

インターネットプロトコルアドレス

コンピュータまたはアプリケーションに割り当てられる数値アドレス。

関連項目 IPエイリアス

インタコネクト (CF)

クラスタインタコネクトを参照。

ウィザード (RMS)

テスト済みのオブジェクト定義を使って特定タイプのアプリケーションを作成するインタラクティブなソフトウェアツール。

ウォームスタンバイ

Oracle Solarisゾーン環境において、運用サーバ、待機サーバともノングローバルゾーンは起動したまま、ノングローバルゾーン内で動作するアプリケーションのみを切り替え、業務を引き継がせる運用。待機のノングローバルゾーンのOSが起動状態となるため、コールドスタンバイより高速な切替えが可能。

ウォッチドックタイマ監視

OSハングやブート異常を監視するタイマ値。

エラー検出 (RMS)

エラーを検出するプロセス。RMSでは、ログの記録開始、ログファイルへのメッセージ送信、リカバリ処理の実行などを行う。

応答待ち時間 (レイテンシ)

データの送信要求を行ってから、実際に応答を受信するまでの時間間隔。

オブジェクト (RMS)

構成定義ファイルまたはシステムグラフでは、ノードは物理または仮想リソースを示す。

関連項目 リーフオブジェクト、オブジェクト定義、ノード状態、オブジェクトタイプ

オブジェクトタイプ (RMS)

ディスクドライブなど監視される同種のリソースをグループ化するカテゴリ。各オブジェクトタイプにはプロパティと呼ばれる固有の属性があり、この属性により実行する監視またはアクションの種類を限定または定義する。リソースを特定のオブジェクトタイプに関連付けると、関連付けたオブジェクトタイプの属性がリソースに適用される。

関連項目 汎用タイプ

オブジェクト定義 (RMS)

RMSの監視対象となるリソースを識別する構成定義ファイルのエントリ。定義された属性により、関連するリソースのプロパティが指定される。オブジェクト定義に関連するキーワードにobjectがある。

関連項目 属性、オブジェクトタイプ

オペレーティングシステム依存 (CF)

オペレーティングシステム本体と、OS非依存のPRIMECLUSTERモジュールとの間のインタフェースを提供するモジュール。

オペレーティングシステム本体

オペレーティングシステムのうち、常にアクティブでシステムコールを実際の処理に変換している部分。

親 (RMS)

1つ以上の子オブジェクトを保持する、構成定義ファイルまたはシステムグラフのオブジェクト。

関連項目 子、構成定義ファイル、システムグラフ

オンラインメンテナンス

ホストのシャットダウンや電源オフの必要なく機器を追加、削除、または交換できる機能。

回線切替装置(Oracle Solaris 10環境のみ)

外部からの回線を複数ノードの間に接続して、RCIにより接続ノードの切替えを行う装置。

下位グループ (GDS)

他のグループに属しているグループ。下位グループにはボリュームを作成できません。

拡張性

作業負荷の増加に動的に対処するコンピューティングシステムの機能。拡張性は、特にインターネットベースのアプリケーションにおいて、インターネットの使用量の増大に伴って重要になる。

カスタムタイプ (RMS)

汎用タイプを参照。

カスタムディテクタ (RMS)

ディテクタを参照。

仮想インタフェース (VIP)

クラスタの複数ノードをシングルシステムイメージとして見せるために、SISが使用する仮想的なIPアドレスまたはノード名。

仮想ディスク

仮想ディスクでは、Solaris論理I/Oシステムの最上位と物理デバイスドライバとの間に擬似デバイスドライバが追加される。擬似デバイスドライバは全ての論理I/O要求を物理ディスク上にマップする。(富士通テクノロジー・ソリューションズ製品から移行のお客様のみ)

関連項目 連結仮想ディスク、ミラー仮想ディスク(VM)、単独仮想ディスク、ストライプ化仮想ディスク

可用性

多くの企業が必要とする、インターネットによる24時間年中無休のアプリケーション稼動環境の達成度を示す指標。実際と計画の使用時間の比較によってシステムの可用性が決まる。

環境変数

グローバルに定義された変数またはパラメタ。

管理LAN

PRIMECLUSTERの構成における、システムコンソールやクラスタ運用管理PCなどが接続されたプライベートローカルエリアネットワーク(LAN)。管理LANには、一般ユーザがアクセスできないため、非常に高いレベルのセキュリティを確保できる。管理LANを使用するかどうかは選択可能。

関連項目 業務LAN

キーワード (予約語)

プログラミング言語において、ある特別な意味を持つ用語。たとえば、構成定義ファイルのnodeキーワードは、後に続く定義の種類を指定する。

キュー

メッセージキューを参照。

業務LAN

一般ユーザがマシンにアクセスするためのローカルエリアネットワーク (LAN)。

関連項目 管理LAN

共用ディスク接続確認

ノード起動時に共用ディスク装置の電源投入漏れやケーブルの結線誤りがないことを確認する機能。

共用リソース

複数ノード間で共有されるディスクドライブなどのリソース。

関連項目 専用リソース、リソース

切替え (RMS)

userApplicationの制御を監視対象の1つのノードから他のノードに切替えるRMSのプロセス。

関連項目 自動切替え、指定切替え、フェイルオーバ、対称切替え

切替方式

GLSが提供するLAN二重化の方式名。高速切替方式、NIC切替方式、GS/SURE連携方式(Solaris)、GS連携方式(Linux)、仮想NIC方式、マルチパス方式(Solaris)が存在する。

クラス (GDS)

ディスククラス (GDS)を参照。

クラスタ

1つのコンピューティングソースに統合されるコンピュータの集まり。クラスタは分散型のパラレルコンピューティングを実行する。

関連項目 RMS構成

クラスタアプリケーション (RMS)

RMSのリソース定義において、userApplicationに分類されるリソース。複数のリソースをアプリケーション単位にグループ化する際に使用される。

クラスタインタコネクト (CF)

PRIMECLUSTERがノード間の通信処理で専用に使用するネットワーク接続。

クラスタ基盤 (CF)

基本OSの上位で動作するPRIMECLUSTERの基本モジュール。PRIMECLUSTERの上位サービスが使用する機能をCF(Cluster Foundation)インタフェースとして提供する。

関連項目 Cluster Foundation

クラスタ構成のバックアップおよびリストア（Solarisのみ）

CCBRを使用すると、あるクラスタノードについて現在のPRIMECLUSTER構成情報を簡単に保存することができる。また、構成情報をリストアすることもできる。

クラスタ参入サービス (CF)

新規クラスタの作成およびクラスタへのノードの追加を処理するPRIMECLUSTERサービス。

クラスタ整合状態 (クォーラム)

クラスタシステムを構成するノード間の整合性が保たれている状態。具体的には、クラスタシステムを構成する、各ノードのCFの状態がUPまたはDOWNである状態 (LEFTCLUSTERとなっているノードが存在しない)。

クラスタパーティション (CF)

クラスタ内の複数ノードのクラスタインタコネクトによる通信が不可能な場合に発生する状態。クラスタクパーティション状態でアプリケーションが共用ディスクにアクセスし続けるとデータの整合性がとれなくなる恐れがある。

関連項目 スプリットブレイン状態

クラスタリソース管理機構

複数のノード間で共用されるハードウェアを管理する機構。

グラフ (RMS)

システムグラフを参照。

グラフィカルユーザインタフェース

ウィンドウ、アイコン、ツールバー、プルダウンメニューを使った、コマンドラインインタフェースより使いやすいコンピュータインタフェース。

グループ (GDS)

ディスクグループ (GDS)を参照。

経路

“PRIMECLUSTERコンセプトガイド” では、ノードとノードの間を接続する冗長化されたクラスタインタコネクトの各々のネットワーク経路を意味している。

ゲートウェイクラスタノード (SIS)

ゲートウェイクラスタノードは外部ネットワークインタフェースを有し、全ての受信パッケージはこのノードで受信され、サービスのスケジューリングアルゴリズムに従って選択したサービスノードに転送される。

関連項目 サービス提供ノード、データベースノード、Scalable Internet Services

子 (RMS)

1つ以上の親に属し、構成定義ファイルに定義されるリソース。子は複数の親に属することが可能。また、子を保持して親ノードとなることも、子を持たずにリーフオブジェクトとなることも可能。

関連項目 リソース、オブジェクト、親、リーフオブジェクト

高可用性

冗長リソースにより一点故障箇所を排除する概念。

構成定義ファイル (RMS)

監視するリソースを定義し、リソース間の相互依存性を設定するRMS構成定義ファイル。デフォルトファイル名はconfig.us。

高速切替方式

GLSが提供するLAN二重化方式の1つ。多重化したLANを同時に使用し、サーバ間通信のスケーラビリティ向上と、LAN異常発生時の高速な切替えを実現する。

コールドスタンバイ

待機状態にあるノードで、すぐに運用状態となるための事前準備を行わない運用。

コンカチネーション

複数の物理ディスクを連結すること。複数のディスクを仮想的に1つの大容量ディスクとして使用する仕組み。

最上位グループ (GDS)

他のグループに属していないグループ。最上位グループには、ボリュームを作成できます。

サービス提供ノード (SIS)

FTP、Telnet、HTTPなど1つ以上のTCPサービスを提供し、ゲートウェイクラスタノードからクライアント要求を受信する。

関連項目 データベースノード、ゲートウェイクラスタノード、Scalable Internet Services

システムグラフ (RMS)

構成定義ファイルの作成、または解釈に使用される監視対象リソースのビジュアル表示 (マップ)。

関連項目 構成定義ファイル

システムディスク (GDS)

動作中のオペレーティングシステムがインストールされているディスク。次のいずれかのファイルシステム (またはスワップ域) として現在動作しているスライスが存在するディスク全体を指します。

Solaris の場合: /、/usr 、/var 、またはスワップ域
Linux の場合: /、/usr 、/var 、/boot、/boot/efi、またはスワップ域

指定切替え (RMS)

管理者がRMSのuserApplicationを指定したノードに切替える処理。

関連項目 自動切替え、フェイルオーバ、切替え、対称切替え

自動切替え (RMS)

ある一定の条件が検出された際に、userApplicationの実行を他のノードへ自動的に切替えるRMSの処理。

関連項目 指定切替え、フェイルオーバ、切替え、対称切替え

自動電源制御

自動電源制御は、ESF (Enhanced Support Facility) で提供している機能で、サーバの電源投入および、切断を自動的に行うための機能である。

シャットダウン機構

異常が発生したノードを強制停止させるための機構。PRIMECLUSTERは、クラスタ整合性 (クォーラム) が保てない状態になったと判断した場合に、シャットダウン機構 (SF) を使用して、クラスタシステムをクラスタ整合状態 (クォーラム) に戻している。

状態

リソース状態を参照。

状態遷移プロシジャ

クラスタ制御からの状態遷移指示を受け取り、リソースの活性/非活性化を制御 (クラスタアプリケーションの起動/停止など) するもの。

冗長化

オブジェクトがクラスタ内の他のオブジェクトのリソース負荷を引継ぐ機能、およびRAIDハードウェア、またはソフトウェアにより2次記憶装置に保存されているデータを複製する機能。

シングルディスク (GDS)

グループに属していないSDXディスクで、シングルボリュームを作成できるディスク。

シングルノードクラスタ

1ノードから構成されるクラスタシステムにおける運用形態。

シングルボリューム (GDS)

グループに属していないシングルディスク内に作成されたボリューム。データは冗長化されません。

スイッチオーバ

ユーザの要求によりユーザ業務が運用系から待機系へ処理やデータを引継ぐこと。

スクリプト (RMS)

リソースの状態遷移に対応してBM (ベースモニタ) から実行されるシェルプログラム。スクリプトによりリソースの状態が変更される場合もある。

スコープ (GDS)

共用タイプのディスククラスにおいてオブジェクトを共用できるノード群の範囲を表します。

ストライピング

データを一定のサイズに分割して、複数のスライスに交互に振り分けて書込むこと。I/Oを複数の物理ディスクに分散して同時に発行する仕組み。

ストライプ化仮想ディスク

ストライプ化仮想ディスクは複数の区画で構成されます。物理パーティションや複数の仮想ディスク (通常はミラーディスク) で構成することもできます。このようにして仮想ディスク上の連続したI/O処理を複数の物理ディスク上のI/O処理に変換することができる。この機能はRAIDレベル0 (RAID0) に該当する。(富士通テクノロジー・ソリューションズ製品から移行のお客様のみ)

関連項目 連結仮想ディスク、ミラー仮想ディスク(VM)、単独仮想ディスク、仮想ディスク

ストライプグループ (GDS)

ストライプ(stripe)タイプのディスクグループ。ストライピングの単位となるディスクおよび下位グループの集まり。

ストライプ幅 (GDS)

ストライピングする際の、データを分割するサイズ。

ストライプボリューム (GDS)

ストライプグループ内に作成されたボリューム。ストライピングによってI/O負荷を複数のディスクに分散させることができます。データは冗長化されません。

スプリットブレイン状態

クラスタパーティションを参照。

スペアディスク (GDS)

故障したディスクの替わりにミラーリング状態を回復させるための予備ディスク。

世代数

PRIMECLUSTER のバックアップ/リストアは、データの世代管理が可能で、現在の世代数は、バックアップおよびリストアデータの名前の一部として付加されます。なお世代数は0以上の整数が使用され、バックアップが成功するたびに1ずつ増加します。世代数は、ccbr.gen ファイル、または、cfbackup(1M)コマンドおよび cfrestore(1M)コマンドのオプション引数にて指定することができます。詳細は、cfbackup(1M)コマンドおよびcfrestore(1M)コマンドのマニュアルページを参照してください。

専用ネットワークアドレス

RFC1918により指定された一定範囲の予約済IPアドレス。どの部門でも使用可能であるが、異なる部門が同時に同じアドレスを使用する可能性があるため、インターネット経由で外部から参照できないようにする必要がある。

専用リソース (RMS)

1台のノードのみが使用可能で、他のRMSノードからは使用できないリソース。

関連項目 リソース、共用リソース

属性 (RMS)

各オブジェクトタイプについて、BM (ベースモニタ) がどう処理するかを規定するオブジェクト。

対称切替え (RMS)

全てのRMSノードが他の任意のRMSノードからリソースを引継ぐことのできる機能。

関連項目 自動切替え、指定切替え、フェイルオーバ、切替え

タイプ

オブジェクトタイプを参照。

多重ホスト

複数のコントローラ経由で同一のディスク。(富士通テクノロジー・ソリューションズ製品から移行のお客様のみ)

単独仮想ディスク

単独仮想ディスクは、物理ディスクパーティションの1領域、またはパーティション全体を定義します。(富士通テクノロジー・ソリューションズ製品から移行のお客様のみ)

関連項目 連結仮想ディスク、ストライプ化仮想ディスク、仮想ディスク

通知メッセージ (RMS)

ディテクタがBM (ベースモニタ) に特定リソースの状態を通知するメッセージ。

停止要求

クラスタ整合状態 (クォーラム) を回復するために、指定したノードを強制停止させるための指示。

ディスククラス (GDS)

SDXオブジェクトの集まり。共用タイプのディスククラスは、PRIMECLUSTERシステムで利用可能なリソースの単位でもあります。ディスククラスを単に「クラス」と呼ぶ場合もあります。

ディスクグループ (GDS)

ミラーリング、ストライピング、またはコンカチネートされる単位となるディスクまたは下位グループの集まり。同じディスクグループに属しているディスクおよび下位グループは、そのディスクグループのタイプ属性(ミラー、ストライプ、またはコンカチネーション)に応じて、互いに ミラーリング、ストライピング、またはコンカチネートされます。 ディスクグループを単に「グループ」と呼ぶ場合もあります。

ディテクタ (RMS)

特定のオブジェクトタイプの状態を監視して、リソースの状態変化をBM (ベースモニタ) に通知するプロセス。

データベースノード (SIS)

SIS構成の設定、動的データ、統計を管理するノード。

関連項目 ゲートウェイクラスタノード、サービス提供ノード、Scalable Internet Services

デーモン

特定の機能を繰り返し実行する、システムに常駐するプロセス。

電源連動 (制御)

クラスタシステムにおいて、1ノードの電源を投入すると、電源切断状態にあるその他全てのノードおよびノードとRCIケーブルで接続されたディスクアレイ装置の電源が投入されること。

ネットワークアダプタ

LAN関連のネットワークアダプタ。

ネットワークインタフェースカード

ネットワークアダプタを参照。

ノード

クラスタのメンバであるホスト。コンピュータノードとはコンピュータのことを指す。

ノード間通信機構

PRIMECLUSTER CFで使用されるクラスタノード間の通信機能。クラスタノード間通信専用に設計されているため、TCP/IPよりもオーバヘッドが少なく、メッセージの到着順も保証したデータグラム通信サービスを行うことができる。

ノード状態 (CF)

クラスタ内の全てのノードは、同じクラスタの他の全てのノードのローカル状態を管理する。クラスタ内のノードは、全てUP、DOWN、またはLEFTCLUSTERのいずれかの状態にある。

関連項目 UP、DOWN、LEFTCLUSTER

パトロール診断

ハードウェアの故障を定期的に診断する機能。

ハブ

LANや、ファイバチャネルで使用されるスター型の結線装置。

汎用タイプ (RMS)

汎用プロパティを持つオブジェクトタイプ。汎用タイプは、既存のオブジェクトタイプに割り当てることのできない監視対象リソースがある場合にRMSをカスタマイズするために使用される。

関連項目 オブジェクトタイプ

非同期監視

SAの機能に加え、リモートクラスタノードの状態を監視し、そのノードのダウンを即時に検出するコンポーネント。

フェイルオーバ (RMS、SIS)

SISでは、このプロセスにより障害発生ノードのバックアップノードへの切替えを行う。RMSでは、このプロセスを切替えと呼ぶ。

関連項目 自動切替え、指定切替え、切替え、対称切替え

フォルトトレラントネットワーク (耐故障性を備えたネットワーク)

耐故障性 (フォルトトレラント) を備えたネットワーク。耐故障性 (フォルトトレラント) とは、コンピュータシステムの一部に何らかの障害が発生した場合でも、正常な動作を保ち続ける能力のこと。よって、フォルトトレラントネットワークとはネットワークシステムの一部に異常が発生した場合でも、正常に通信を継続できるネットワークのことを意味している。

物理IPアドレス

ネットワークインタフェースカードのインタフェース (たとえばhme0) に直接割り振られたIPアドレス。

プライマリノード (RMS)

RMSの起動時にユーザアプリケーションをオンラインにするデフォルトノード。userApplicationのオブジェクト定義中に最初に記述されたノードがプライマリノードとなる。

ホットスタンバイ

待機状態にあるノードで、すぐに運用状態となるための事前準備を行う運用。

ボリューム (GDS)

論理ボリューム (GDS)を参照。

マウントポイント

ディレクトリツリー上でファイルシステムが接続されるポイント。

ミラー仮想ディスク(VM)

ミラー仮想ディスクは複数の物理デバイスで構成され、全ての出力処理が全てのデバイス上で同時実行される。(富士通テクノロジー・ソリューションズ製品から移行のお客様のみ)

関連項目 連結仮想ディスク、単独仮想ディスク、ストライプ化仮想ディスク、仮想ディスク

ミラーグループ (GDS)

ミラー(mirror)タイプのディスクグループ。互いにミラーリングされるディスクおよび下位グループの集まり。

ミラーボリューム (GDS)

ミラーグループ内に作成されたボリューム。ミラーリングによってデータが冗長化されます。

ミラーリング

同じデータを複数のスライスに書込むことによって、冗長性を維持すること。一部のスライスで障害が発生したとしても、正常なスライスが残っていれば、ボリュームへのアクセスが継続できる仕組み。

メタデータサーバ

ファイルシステム (メタデータ) の制御情報を一括管理するGFSデーモン。

メッセージ

1つのソフトウェアプロセスから他のプロセス、デバイス、またはファイルに伝送されるデータの集まり。

メッセージキュー

メッセージの保存場所として使用される専用のメモリ領域。

モデル混在クラスタ

SPARC Enterprise の異なるモデルによって構築したクラスタシステム。たとえば1つのノードがSPARC Enterprise M3000 で、もう1つのノードが SPARC Enterprise M4000など。
モデルは、代表的なマシンではSPARC M12-2/M12-2S、SPARC M10-1/M10-4/M10-4S、SPARC S7-2/S7-2L、SPARC T7-1/T7-2/T7-4、SPARC T5-2/T5-4/T5-8、SPARC T4-1/T4-2/T4-4、SPARC T3-1/T3-2/T3-4、SPARC Enterprise T1000/T2000、SPARC Enterprise T5120/T5220/T5140/T5240/T5440、SPARC Enterprise M3000/M4000/M5000/M8000/M9000 で分かれます。

ユーザグループ

Web-Based Admin ViewやCluster Admin GUIが提供する環境設定、運用管理などの操作範囲を限定するもので、wvroot、clroot、cladmin、clmonの4種類がある。クラスタ管理サーバのオペレーションシステムの管理者に依頼して、個々のユーザIDを適切なユーザグループへ登録する。

リーフオブジェクト (RMS)

システムグラフの最下位オブジェクト。リーフオブジェクトは構成定義ファイルの最後に定義される。リーフオブジェクトはその配下に子オブジェクトを持たない。

リソース (RMS)

ミラーディスク、ミラーディスク部品、データベースサーバなどの機能を提供する、専用または共用のハードウェアまたはソフトウェア要素。ローカルリソースは、ローカルノード上でのみ監視対象となる。

関連項目 専用リソース、共用リソース

リソース状態 (RMS)

リソースの現在の状態。

リソース定義 (RMS)

オブジェクト定義を参照。

リソースデータベース

複数のノード間で共用されるハードウェアの情報を管理するデータベース。リソースデータベースは、クラスタリソース管理機構により管理される。

リソースラベル (RMS)

システムグラフに表示されるリソース名。

リモートコンソール接続装置

RS232CインタフェースとLANインタフェースを変換する装置。本装置により、LAN接続された他の装置 (パソコン) からTelnet機能によりTTYコンソール機能を利用可能とする。

リモートノード

リモートホストを参照。

リモートホスト

遠距離通信回線またはLANを使ってアクセスするホスト。

関連項目 ローカルホスト

リンク (RMS)

特定リソース間の親子関係を指定する。

連結仮想ディスク

1つ以上のディスクドライブ上の複数の区画で構成され、各部を合計したものに相当する。ディスクを細かく分割する単独仮想ディスクと異なり、各ディスクまたはパーティションを連結して1つの大規模な論理ディスクを構成する。(富士通テクノロジー・ソリューションズ製品から移行のお客様のみ)

関連項目 ミラー仮想ディスク(VM)、単独仮想ディスク、ストライプ化仮想ディスク、仮想ディスク

ローカルMACアドレス

ローカルエリアネットワーク (LAN) のシステムごとに、システム管理者がそのシステム内部での一意性を保証するMACアドレス。

ローカルエリアネットワーク

業務LANを参照。

ローカルホスト

コマンドまたはプロセスを開始するホスト

関連項目 リモートホスト

ログファイル

重要なシステムイベントやメッセージを記録したファイル。BM (ベースモニタ)、ウィザード、ディテクタにはそれぞれ固有のログファイルがある。

ローリングアップデート

クラスタシステムにおいて、修正適用、保守時に使用されるアップデート手法。1ノードずつ順次修正適用を行うことで、業務を停止せずに修正を適用することが可能となる。

論理ボリューム (GDS)

利用者が直接アクセスできる仮想ディスクデバイスの総称。利用者は、物理ディスクのスライス
(パーティション)と同じように、論理ボリュームにアクセスできます。論理ボリュームを単に「ボリューム」と呼ぶ場合もあります。

索 引

[A]

Attribute....46,48

[B]

BM....47

[C]

CF....23,24,43
CF/IP....40
CIM....4
Cluster Admin....23,24
Cluster Admin GUI....50
Cluster Foundation....23,43

[D]

Data Link Provider Interface....44
DLPI....44

[E]

eXtended System Control Facility....30

[G]

GDS....23,32
GFS....23,34
GFS共用ファイルシステム....34
GLS....24,36
GS/SURE連携方式（Solaris）、GS連携方式(Linux)....39
GUI....24

[H]

HAマネージャ....2

[I]

ICF....40
IPエイリアス....47,52
iRMC非同期監視....30

[K]

K5環境....9
KVM環境....8

[M]

MA....25
MMB非同期監視....30
Monitoring Agent....5

[N]

NIC切替方式....37

[O]

Object Definition....48
Object Type....48
Oracle Solarisカーネルゾーン環境....10
Oracle Solarisノングローバルゾーン環境....11
Oracle VM Server for SPARC環境....9

[P]

Parallel Application Services....23,32
PAS....23,32
PRIMECLUSTER SF....4,23,25
PRIMECLUSTERのモジュール....23
PRIMEQUEST 2000....30
PRIMEQUEST 3000....30

[R]

RAID....64
RCI非同期監視....28
Reliant Monitor Services....23
RHOSP環境....8
RMS....23,31,45
RMS Wizard Tools....5,32,47,51
RMSウィザード....23,32,46,51
RMS環境変数....50
RMS構成....51
RMSの監視と切替え....47
RMSの管理....50
RMSリソース....46,48,49
RMSリソース名....48

[S]

SA....25
SAN....33,52
SD....25
SNMP非同期監視....30
SPARC Enterprise Mシリーズ....28
Storage Area Network....33,52
System Control Facility....28

[V]

VMware環境....9

[W]

Web-Based Admin View....23,24

[あ]

アプリケーションの切替え....46
アーキテクチャ....21
インタコネクトプロトコル....40
ウィザード....5
エラー率....44
応答待ち時間....43
オブジェクト....46
オブジェクト間の依存関係....47
オブジェクトタイプ....48
オブジェクト定義....48

[か]

拡張性....1,6,23
カスタマイズオプション....50
仮想NIC方式（Linux）....38
仮想NIC方式（Solaris）....38
仮想化対応....7
仮想ディスク....52
可用性....23
擬似デバイスドライバ....60

共用ディスク装置 52
切替処理 46
クラスタインタコネクト 2,40
クラスタシステムの運用 24
クラスタシステムの管理 24
クラスタシステムの構築 24
クラスタ整合性モニタ 4
クラスタパーティション 2
クラスタリング 1
経路 41
高可用性 1,2,45
高信頼化 36
構成スクリプト 49
構成定義ファイル 46,47,48
高速切替方式 37
コンソール非同期監視 28

[さ]

自動切替え 46
シャットダウンエージェント 25
シャットダウン機構 4,25
シャットダウンデーモン 25
手動切替え 46
障害の検出 50
状態トリガ・スクリプト 49
状態変更 47
冗長化 36,41,45
冗長手法 45
シングルノードクラスタ 1,6
診断機構 24
信頼性 44
スケーラビリティ 6,23
セキュリティ 44
設計時の検討項目 42
属性 46,48

[た]

帯域幅 42
ディテクタ 49,50
デバイスインタフェース 44
伝送路二重化機能 36
データ整合性の保証 2
データの整合性 47
データの整合性保証 23

[な]

ノード間通信機構 40
ノード間のデータ引継ぎ 6

[は]

パトロール診断機能 6
汎用リソースタイプ 50
ハートビート 40,42,43
非同期監視 5,25,27
フェイルオーバ 46
プラットフォーム非依存性 22
プロセス監視機構 32
ベースモニタ 47
ボリューム管理機能 32

[ま]

マルチパス機能 37
ミラーリング機能 32
モジュール方式 22

[や]

ユーザ業務の監視 2
ユーザ業務の起動 51
ユーザ定義スクリプト 46
ユーティリティウィザード 52
要求トリガ・スクリプト 49

[ら]

リカバリ方法 50
リソースの状態 49
レイテンシ 43