Sun Cluster の概要 (Solaris OS 版)

Sun Microsystems, Inc.
4150 Network Circle
Santa Clara, CA 95054
U.S.A.

Part No: 820–6910–10
2009 年 1 月、 Revision A

090219@21990

目次

はじめに

Sun™Cluster Overview for Solaris OS では、Sun Cluster 製品を紹介します。製品の目的および Sun Cluster がこの目的を実現するために利用できる方法が記載されています。本書では、Sun Cluster の主要な概念についても説明します。読者は、本書を通して Sun Cluster の特長や機能を知ることができます。

関連マニュアル

関連する Sun Cluster トピックについての情報は、以下の表に示すマニュアルを参照してください。Sun Cluster のドキュメントはすべて http://docs.sun.com から利用できます。

項目	マニュアル
概要	『Sun Cluster の概要 (Solaris OS 版)』
	『Sun Cluster 3.2 1/09 Documentation Center 』
概念	『Sun Cluster の概念 (Solaris OS 版)』
ハードウェアの設計と管理	『Sun Cluster 3.1 - 3.2 Hardware Administration Manual for Solaris OS 』
	各ハードウェア管理ガイド
ソフトウェアのインストール	『Sun Cluster ソフトウェアのインストール (Solaris OS 版)』
	『Sun Cluster クイックスタートガイド (Solaris OS 版)』
データサービスのインストールと管理	『Sun Cluster データサービスの計画と管理 (Solaris OS 版)』
	各データサービスガイド
データサービスの開発	『Sun Cluster データサービス開発ガイド (Solaris OS 版)』
システム管理	『Sun Cluster のシステム管理 (Solaris OS 版)』
	『Sun Cluster Quick Reference 』
ソフトウェアアップグレード	『Sun Cluster Upgrade Guide for Solaris OS 』

項目	マニュアル
エラーメッセージ	『Sun Cluster Error Messages Guide for Solaris OS 』
コマンドと関数のリファレンス	『Sun Cluster Reference Manual for Solaris OS 』
	『Sun Cluster Data Services Reference Manual for Solaris OS 』
	『Sun Cluster Quorum Server Reference Manual for Solaris OS 』

Sun Cluster ドキュメントの完全なリストについては 、http://wikis.sun.com/display/SunCluster/Home/ で Sun Cluster ソフトウェアの使用しているリリースのリリースノートを参照してください。

マニュアル、サポート、およびトレーニング

Sun の Web サイトでは、次のサービスに関する情報も提供しています。

- マニュアル (http://jp.sun.com/documentation/)
- サポート (http://jp.sun.com/support/)
- トレーニング (http://jp.sun.com/training/)

問い合わせについて

Sun Cluster システムのインストールや使用に関して問題がある場合は、以下の情報をご用意の上、担当のサービスプロバイダにお問い合わせください。

- 名前と電子メールアドレス (利用している場合)
- 会社名、住所、および電話番号
- システムのモデルとシリアル番号
- オペレーティング環境のリリース番号 (例: Solaris 9)
- Sun Cluster ソフトウェアのバージョン番号 (例: 3.2 1/09)

次のコマンドを使用し、システム上の各 Solaris ホストに関して、サービスプロバイダに必要な情報を収集してください。

コマンド	機能
`prtconf -v`	システムメモリのサイズと周辺デバイス情報を表示します
`psrinfo -v`	プロセッサの情報を表示する
`showrev -p`	インストールされているパッチを報告する
`prtdiag -v`	システム診断情報を表示する

コマンド	機能
`scinstall -pv`	Sun Cluster ソフトウェアのリリースおよびパッケージのバージョン情報を表示する
`scstat`	クラスタの状態のスナップショットを提供します
`scconf -p`	クラスタ構成情報を表示します
`scrgadm -p`	インストールされているリソースやリソースグループ、リソースタイプの情報を表示する

上記の情報にあわせて、`/var/adm/messages` ファイルの内容もご購入先にお知らせください。

表記上の規則

このマニュアルでは、次のような字体や記号を特別な意味を持つものとして使用します。

表 P–1 表記上の規則

字体または記号	意味	例
`AaBbCc123`	コマンド名、ファイル名、ディレクトリ名、画面上のコンピュータ出力、コード例を示します。	`.login` ファイルを編集します。 `ls -a` を使用してすべてのファイルを表示します。 `system%`
`AaBbCc123`	ユーザーが入力する文字を、画面上のコンピュータ出力と区別して示します。	`system%` **`su`** `password:`
AaBbCc123	変数を示します。実際に使用する特定の名前または値で置き換えます。	ファイルを削除するには、`rm` *filename* と入力します。
『』	参照する書名を示します。	『コードマネージャ・ユーザーズガイド』を参照してください。
「」	参照する章、節、ボタンやメニュー名、強調する単語を示します。	第 5 章「衝突の回避」を参照してください。 この操作ができるのは、「スーパーユーザー」だけです。

表 P–1 表記上の規則 (続き)

字体または記号	意味	例
\	枠で囲まれたコード例で、テキストがページ行幅を超える場合に、継続を示します。	`sun% grep '^#define \` `XV_VERSION_STRING'`

コード例は次のように表示されます。

- C シェル

```
machine_name% command y|n [filename]
```

- C シェルのスーパーユーザー

```
machine_name# command y|n [filename]
```

- Bourne シェルおよび Korn シェル

```
$ command y|n [filename]
```

- Bourne シェルおよび Korn シェルのスーパーユーザー

```
# command y|n [filename]
```

[ ] は省略可能な項目を示します。上記の例は、*filename* は省略してもよいことを示しています。

| は区切り文字 (セパレータ) です。この文字で分割されている引数のうち 1 つだけを指定します。

キーボードのキー名は英文で、頭文字を大文字で示します (例: Shift キーを押します)。ただし、キーボードによっては Enter キーが Return キーの動作をします。

ダッシュ (-) は 2 つのキーを同時に押すことを示します。たとえば、Ctrl-D は Control キーを押したまま D キーを押すことを意味します。

◆ ◆ ◆ 第 1 章

Sun Clusterの概要

Sun Cluster 構成はハードウェアと Sun Cluster ソフトウェアが統合されたソリューションであり、高度な可用性とスケーラビリティーを備えたサービスを提供するために使用されます。 この章では、Sun Cluster 機能の概要を説明します。

この章で説明する内容は次のとおりです。

Sun Cluster によるアプリケーションの可用性の向上

クラスタとは、緩やかに結合された処理ノードの集合のことで、データベース、Web サービス、ファイルサービスなどのネットワークサービスやアプリケーションを、クライアントからは 1 つのシステムに見える形で提供します。

クラスタ環境では、すべてのノードがインターコネクトによって接続され、単一のエンティティーとして動作するので、可用性と性能が向上します。

HA を備えたクラスタは、通常、単一のサーバーシステムなら停止するような障害が発生しても、データやアプリケーションに対してほとんど連続的なアクセスを提供するように稼動し続けることができます。ハードウェア、ソフトウェア、またはネットワークの単一の故障によりクラスタに障害が発生することはありません。これに対して、フォルトトレラントのハードウェアシステムは、データとアプリケーションに対する一定したアクセスを可能にしますが、特殊なハードウェアが必要なため、コストが高くなります。フォルトトレラントシステムには通常、ソフトウェア障害に対する備えはありません。

個々の Sun Cluster システムは密接に関わり合ったノードの集合であり、すべてのネットワークサービスやアプリケーションが一元的に管理されます。Sun Cluster システムは、次のハードウェアとソフトウェアの組み合わせを通して HA を実現します。

- 冗長化されたディスクシステム。ストレージを提供するこれらのディスクシステムは一般にミラー化されるため、ディスクやサブシステムに障害が発生しても、操作が中断されることはありません。さらに、ディスクシステムへの接続は冗長化されているため、サーバーやコントローラ、ケーブルに障害が発生しても、データにアクセスできなくなることはありません。リソースへのアクセスは、Solaris ホスト間を結ぶ高速インターコネクトを通して行われます。さらに、クラスタのすべてのホストがパブリックネットワークに接続されているため、複数のネットワークに散在するクライアントからクラスタにアクセスできます。
- 電源装置や冷却システムなど、冗長化されたホットスワップ可能コンポーネント。これらのコンポーネントは冗長化されているため、ハードウェアに障害が発生しても、システムは操作を続けることができ、可用性が向上します。ハードウェアコンポーネントがホットスワップ可能であれば、そのコンポーネントを動作中のシステムから取り外したり、システムに追加することができます。そのためにシステムを停止する必要はありません。
- Sun Cluster ソフトウェアフレームワーク。このフレームワークはノードの障害を素早く検知し、それと同一環境で動作する別のノードにアプリケーションやサービスを移行します。すべてのアプリケーションが同時に使用不能になることはありません。停止したノードと関係のないアプリケーションは、この復旧処理の間も全面的に使用可能です。さらに、障害が発生したノードのアプリケーションは、復旧されると同時に使用可能になります。復旧したアプリケーションは、ほかのすべてのアプリケーションが完全に復旧するまで待つ必要はありません。

可用性の管理

システムで単一ソフトウェアまたはハードウェアの障害が発生してもあるアプリケーションが稼働し続けられる場合、そのアプリケーションには高い可用性があります。ただし、アプリケーション自体のバグやデータ破損に起因する障害の場合は除きます。HA のアプリケーションには次が適用されます。

- リソースを使用するアプリケーションから、復旧は透過的に行われます。
- リソースのアクセスは、ノードに障害が発生しても完全に保持されます。
- アプリケーションのホストノードが別のノードに移行されたことをアプリケーションが検知することはありません。
- 単一ノードの障害は、このノードに接続されているファイルやデバイス、ディスクボリュームを使用する、その他の障害を受けないノード上のプログラムに対し、完全に透過的です。

フェイルオーバーサービスとスケーラブルサービス、およびパラレルアプリケーション

フェイルオーバーサービスやスケーラブルサービス、パラレルアプリケーションを使用すると、アプリケーションの高い可用性が実現し、クラスタで動作するアプリケーションの性能が向上します。

フェイルオーバーサービスでは、冗長性を通して HA を提供します。障害が発生した場合、ユーザーが介入することなく、アプリケーションの設定に従って、稼動しているアプリケーションを同じノードで再起動するか、クラスタの別のノードに移動することができます。

スケーラブルサービスでは、性能を高めるために、クラスタの複数のノードでアプリケーションを同時に実行します。スケーラブルな構成では、クラスタ内の各ノードが、データを提供して、クライアント要求を処理することができます。

PDB (パラレルデータベース) を使用すれば、データベースサーバーの複数のインスタンスを使って次のことができます。

- クラスタに参加する。
- 同じデータベースに対する別々のクエリーを同時に処理する。
- 大規模なクエリーの場合、クエリーを並列に処理する。

フェイルオーバーサービスとスケーラブルサービス、およびパラレルアプリケーションの詳細については、30 ページの「データサービスのタイプ」を参照してください。

IP ネットワークマルチパス

クライアントは、パブリックネットワークを介してクラスタにデータ要求を行います。各 Solaris ホストは、1 つまたは複数のパブリックネットワークアダプタを介して少なくとも 1 つのパブリックネットワークに接続されています。

IP ネットワークマルチパス では、サーバーの複数のネットワークポートを同じサブネットに接続できます。IP ネットワークマルチパス ソフトウェアはネットワークアダプタ障害からの復旧をサポートします。そのために、まず、ネットワークアダプタの障害や修復を検知し、次に、アダプタと代替アダプタとの間でネットワークアドレスを同時に切り替えます。複数のネットワークアダプタが機能している場合、IP ネットワークマルチパス は、送信パケットをアダプタ間に分配することによってデータスループットの向上を図ります。

記憶装置の管理

多重ホストストレージではディスクが複数の Solaris ホストに接続されるため、ディスクの高い可用性が実現されます。この場合、データに複数のパスを通してアクセスできるため、1 つのパスに障害が発生しても、別のホストがその代わりをします。

多重ホストディスクの使用によって、次のクラスタ処理が可能になります。

- 単一ホストに障害が発生しても処理を継続する。
- アプリケーションデータやアプリケーションバイナリ、構成ファイルを一元化する。
- ホストの障害からユーザーを保護する。クライアント要求があるホストを介してデータにアクセスしているときにそのホストに障害が発生した場合、これらの要求は、同じディスクに対する直接接続を持つ別のホストを使用するようにスイッチオーバーされます。
- ディスクを「マスター」する主ホストを通したグローバルなアクセス、またはローカルパスを通した直接かつ並列のアクセスを提供する。

ボリューム管理のサポート

ボリュームマネージャーを使用すると、大量のディスクやそこに格納されているデータを管理することができます。ボリュームマネージャーは、次のような機能を使ってストレージの容量やデータの可用性を高めます。

- ディスクドライブのストライピングやコンカチネーション
- ディスクのミラー化
- ディスクドライブのホットスペア
- ディスク障害への対応とディスクの交換

Sun Cluster システムは、次のボリュームマネージャーをサポートします。

- Solaris Volume Manager
- Solaris Volume Manager for Sun Cluster (Oban)
- Veritas Volume Manager

Solaris I/O マルチパス (MPxIO)

Solaris I/O マルチパス (MPxIO) (以前の名称は Sun StorEdge Traffic Manager) は、Solaris オペレーティングシステム I/O フレームワークに完全に統合されています。Solaris I/O マルチパスを使用すると、Solaris オペレーティングシステムの単一インスタンス内にある複数の I/O コントローラインタフェースを通してアクセス可能なデバイスを、表示および管理することができます。

Solaris I/O マルチパスアーキテクチャーは、次の機能を提供します。

- 入出力コントローラの障害による入出力の中断を防止する。
- 入出力コントローラの障害時に代替のコントローラに自動的に切り替える。
- 複数の入出力チャネルに負荷をロードバランスさせることによって、入出力の性能を高める。

ハードウェア独立ディスク冗長アレイサポート

Sun Cluster システムでは、ハードウェア独立ディスク冗長アレイ (Redundant Array of Independent Disks、RAID) やホストベースのソフトウェア RAID が使用できます。ハードウェア RAID では、ストレージアレイまたはストレージシステムのハードウェアの冗長性を使って、個々のハードウェア障害がデータの可用性に影響がないようにします。別々のストレージアレイ間でデータがミラー化されている場合には、ホストベースの RAID を使って、個別のハードウェア障害 (ある1つのストレージアレイが完全にオフライン) がデータの可用性に影響がないようにします。ハードウェア RAID とホストベースのソフトウェア RAID を同時に使用することもできますが、ある程度の高いデータ可用性を維持するために、1 つの RAID ソリューションだけを使用することもできます。

クラスタファイルシステムのサポート

クラスタシステム本来の特性の 1 つにリソースの共有があります。そのため、クラスタには、ファイルを一貫性のある方法で共有できるファイルシステムが欠かせません。Sun Cluster のファイルシステムでは、クラスタファイルシステムにより、ユーザーやアプリケーションはリモートまたはローカルの標準 UNIX API を使用して、任意のノードの任意のファイルにアクセスできます。Sun Cluster システムは、次のクラスタファイルシステムをサポートします。

- Solaris ZFS™
- UNIX ファイルシステム (UFS)
- Sun StorEdge QFS ファイルシステム、および Sun QFS 共有ファイルシステム
- Sun Cluster プロキシファイルシステム (PxFS)
- VERITAS ファイルシステム (VxFS)

アプリケーションが、あるノードから別のノードに移動されても、そのアプリケーションは変更なしで同じファイルにアクセスできます。さらに、既存のアプリケーションでクラスタファイルシステムを使用する場合、アプリケーションを変更する必要はありません。

構内クラスタ

標準の Sun Cluster システムは、高可用性と信頼性を 1 箇所から集中的に実現します。地震、洪水、停電などの予測不可能な災害の発生後でもアプリケーションを使用可能なまま維持する必要がある場合は、クラスタを構内クラスタとして構成できます。

構内クラスタでは、数キロメートル離れた別の建物に Solaris ホストや共有ストレージなどのクラスタコンポーネントを配置できます。Solaris ホストと共有ストレージを分離し、それらを企業構内の別の場所や、数キロメートルの範囲にある別の施設内に配置することが可能です。1 箇所に災害が発生しても、残存するホストが障害の発生したホストのサービスを引き継ぐことができます。これにより、ユーザーは引き続きアプリケーションとデータを使用できます。構内クラスタの構成についての詳細は、『Sun Cluster 3.1 - 3.2 Hardware Administration Manual for Solaris OS』を参照してください。

障害の監視

Sun Cluster システムでは、多重ホストディスク、マルチパス、およびクラスタファイルシステムを使って、ユーザーとデータ間のパスの高い可用性を維持します。Sun Cluster システムは、次のコンポーネントの障害を監視します。

- アプリケーション – ほとんどの Sun Cluster データサービスは、データサービスの健全性を周期的に検証するフォルトモニターを備えています。フォルトモニターは、アプリケーションデーモンが動作しているかどうかや、クライアントにサービスが提供されているかどうかを検証します。さらに、フォルトモニターは、検証機能から返される情報に基づいて、デーモンの再起動やフェイルオーバーの指示など、事前に定義されたアクションを開始できます。
- ディスクパス – Sun Cluster ソフトウェアは、ディスクパス監視機能 (DPM) をサポートします。DPM は二次ディスクパスの障害を報告することによって、フェイルオーバーやスイッチオーバーの信頼性を全体的に向上します。
- インターネットプロトコル (IP) マルチパス – Sun Cluster システムで動作する Solaris IP ネットワークマルチパス (IPMP) ソフトウェアは、パブリックネットワークアダプタを監視する基本的なメカニズムです。さらに、障害が検知されると、IPMPソフトウェアは、IP アドレスをあるアダプタから別のアダプタにフェイルオーバーします。
- 定足数デバイス – Sun Cluster ソフトウェアは、定足数デバイス監視機能をサポートしています。定足数デバイス上で定足数が動作しているかどうかが周期的にテストされます。Sun Cluster ソフトウェアは、障害を検出すると、障害を報告して、正常に動作していない定足数デバイスをマーク付けします。以前は障害を起こしていた定足数デバイスが正常な処理に戻っているのを発見すると、自動的にその定足数デバイスをサービスに復帰させます。定足数デバイスをサービスに復帰

させるときには、デバイスに正しい定足数予約情報が配置されます。Sun Cluster システムは、保守モードでない構成済みの定足数デバイスを、種類にかかわらず、自動的に監視します。

管理と構成のためのツール

Sun Cluster システムのインストールや構成、管理は、Sun Cluster Manager GUI またはコマンド行インタフェース (Command-Line Interface、CLI) を使って行うことができます。

さらに、Sun Cluster システムには、Sun Management Center ソフトウェアの中で動作するモジュールが含まれています。これは、クラスタの一部の作業を行う時の GUI になります。

Sun Cluster Manager

Sun Cluster Manager は、Sun Cluster システムの管理に使用するブラウザベースのツールです。管理者は、Sun Cluster Manager ソフトウェアを通して、システムの管理や監視、ソフトウェアのインストール、システムの構成を行うことができます。

Sun Cluster Manager ソフトウェアには、次の機能があります。

- 組み込まれたセキュリティーや認証のメカニズム
- Secure Sockets Layer (SSL) のサポート
- 役割によるアクセス制御 (Role-Based Access Control、RBAC)
- Pluggable Authentication Module (PAM)
- NAFO および IPMP グループ管理機能
- 定足数デバイスやトランスポート、共有ストレージデバイス、リソースグループの管理
- プライベートインターコネクトの高度なエラーチェックや自動検知

コマンド行インタフェース

Sun Cluster コマンド行インタフェース (Command-Line Interface、CLI) は、Sun Cluster システムのインストールや管理を行なったり、Sun Cluster ソフトウェアのボリュームマネージャー部分を管理する一連のユーティリティーです。

Sun Cluster CLI では次の Sun Cluster の管理作業を実行できます。

- Sun Cluster 構成の確認
- Sun Cluster ソフトウェアのインストールと構成
- Sun Cluster 構成の更新

- リソースタイプの登録、リソースグループの作成、リソースグループ内のリソースの起動を管理する
- リソースグループとデバイスグループのノードのマスターや状態を変更する
- 役割によるアクセス制御 (Role-Based Access Control、RBAC) に基づくアクセス制御
- クラスタ全体の停止

Sun Management Center

Sun Cluster システムには、Sun Management Center ソフトウェアの中で動作するモジュールが含まれています。Sun Management Center ソフトウェアは、管理や監視の操作を行う際のクラスタの基盤となるものです。システム管理者は、GUI や CLI を通して次の作業を行うことができます。

- 遠隔システムの構成
- 性能の監視
- ハードウェアやソフトウェア障害の検知と分離

Sun Management Center ソフトウェアは、Sun Cluster サーバー内での動的再構成を管理するインタフェースとしても使用されます。動的再構成には、ドメインの作成や、ボードの動的な接続、動的な切り離しがあります。

役割によるアクセス制御

従来の UNIX システムでは、root ユーザー (スーパーユーザー) はすべての権限を持ちます。つまり、任意のファイルに対する読み取り権と書き込み権、すべてのプログラムの実行権、および任意のプロセスに終了シグナルを送信する権限があります。Solaris の役割によるアクセス制御 (Role-Based Access Control、RBAC) は、権限をすべて与えるかまったく与えないかの二者択一的なスーパーユーザーモデルに代わるものです。RBAC では、基本的に最小限の特権以外は許可しません。つまり、そのユーザーに必要な特権だけを許可します。

RBAC を使用すれば、スーパーユーザーの権限を分割し、それらの権限を特別なユーザーアカウントや役割としてパッケージ化し、それによって、権限を特定の個人に割り当てることができます。このような分割やパッケージ化によって、さまざまなセキュリティーポリシーの作成が可能になります。たとえば、セキュリティーやネットワーキング、ファイアウォール、バックアップ、システム操作など、さまざまな分野で特定目的の管理者用アカウントを設定できます。

◆ ◆ ◆ 第 2 章

Sun Clusterの主要な概念

この章では、Sun Clusterシステムのハードウェアやソフトウェアに関連する主な概念を説明します。ユーザーは、Sun Clusterシステムを使用する前にこれらの概念を理解しておく必要があります。

この章で説明する内容は次のとおりです。

クラスタ、ノード、およびホスト

クラスタとは、1つまたは複数のノードの集合をいい、それらのノードはその集合に排他的に属します。Solaris 10 OS上で実行されるクラスタの種類には、グローバルクラスタとゾーンクラスタがあります。Solaris 10 OSより前にリリースされたいずれかのバージョンのSolaris OS上で実行されるクラスタでは、ノードとは、クラスタメンバーを構成するが定足数デバイスではない物理マシンをいいます。Solaris 10 OS上で実行されるクラスタでは、ノードの概念が変更されています。ノードとは、クラス

タに関連付けられている Solaris のゾーンです。この環境では、*Solaris* ホスト (または単にホスト) とは、Solaris OS およびそのプロセスを実行する、次のハードウェア/ソフトウェア構成のいずれかを指します。

- 仮想マシンまたはハードウェアドメインとして構成されていない、「ベアメタル」物理マシン
- Sun Logical Domains (LDoms) のゲストドメイン
- Sun Logical Domains (LDoms) の I/O ドメイン
- ハードウェアドメイン

Solaris 10 環境では、投票ノードとは、定足数投票、つまりクラスタのメンバーシップ投票の、総数の票を構成するゾーンです。この総数により、そのクラスタが処理を継続するのに十分な票を持っているかどうかが決定されます。非投票ノードとは、定足数投票、つまりクラスタのメンバーシップ投票の、総数を構成しないゾーンです。

クラスタ環境では、すべてのノードがインターコネクトによって接続され、単一のエンティティーとして動作するので、可用性と性能が向上します。

Solaris 10 環境では、グローバルクラスタは、1 つまたは複数のグローバルクラスタ投票ノード、およびオプションで 0 個以上のグローバルクラスタ非投票ノードから構成されるクラスタです。

注 – グローバルクラスタには、オプションで `solaris8`、`solaris9`、`lx` (linux)、またはネイティブブランドの非大域ゾーンを含めることができます。これらはノードではなく、高可用性のコンテナ (リソース) です。

グローバルクラスタ投票ノードとは、グローバルクラスタ内のネイティブブランドの大域ゾーンで、定足数投票、つまりクラスタのメンバーシップ投票の、総数の票を構成します。この総数により、そのクラスタが処理を継続するのに十分な票を持っているかどうかが決定されます。グローバルクラスタ非投票ノードとは、グローバルクラスタ内のネイティブブランドの非大域ゾーンで、定足数投票、つまりクラスタのメンバーシップ投票の、総数の票を構成しません。

Solaris 10 環境では、ゾーンクラスタは、1 つまたは複数のクラスタブランドの投票ノードのみから構成されるクラスタです。ゾーンクラスタは、グローバルクラスタに依存しており、したがって、グローバルクラスタを必要とします。グローバルクラスタはゾーンクラスタを含みません。ゾーンクラスタを構成するには、グローバルクラスタが必要です。ゾーンクラスタは 1 つのマシン上に最大で 1 つのゾーンクラスタノードを持ちます。

注 – ゾーンクラスタノードは、同一マシン上のグローバルクラスタ投票ノードが処理を継続している間にかぎり、処理を継続できます。あるマシン上のグローバルクラスタ投票ノードに障害が発生すると、そのマシン上のすべてのゾーンクラスタノードにも同様に障害が発生します。

Sun Cluster ソフトウェアは、ハードウェア構成に応じて、1 つのクラスタで 1 - 16 個の Solaris ホストを使用できます。使用しているハードウェア構成でサポートされる Solaris ホストの数については、ご購入先にお問い合わせください。

1 つのクラスタ内の Solaris ホストは、通常、1 つ以上のディスクに接続されます。ディスクに接続されていない Solaris ホストは、クラスタファイルシステムを使用して多重ホストディスクにアクセスします。並列データベース構成の下にある各 Solaris ホストは、一部またはすべてのディスクに同時にアクセスします。

クラスタ内のすべてのノードは、別のノードがいつクラスタに結合されたか、またはクラスタから切り離されたかを認識します。さらに、クラスタ内のすべてのノードは、ローカルに実行されているリソースだけでなく、他のクラスタノードで実行されているリソースも認識します。

同じクラスタ内の各 Solaris ホストの処理、メモリー、および入出力機能を同等にして、フェイルオーバー発生時にパフォーマンスが著しく低下しないようにする必要があります。フェイルオーバーの可能性があるため、各ホストには、ノードの障害時にもサービスレベル合意を満たせる十分な容量を持たせる必要があります。

ゾーンクラスタ

この節では、ゾーンクラスタの主要な機能および利点について説明します。

ゾーンクラスタの機能および利点

ゾーンクラスタを使用することで、次のような機能および利点が得られます。

- アプリケーションの障害分離 – あるゾーンクラスタでアプリケーションに障害が発生しても、その他のゾーンクラスタのアプリケーションには影響しません。たとえば、あるゾーンクラスタのノードが起動、停止、または再起動しても、その他のゾーンクラスタは影響を受けません。
- セキュリティー – あるゾーンクラスタノードにログインしているアプリケーションまたはユーザーは、グローバルクラスタまたはその他のゾーンクラスタの要素を参照したり変更したりすることはできません。ゾーンクラスタには、そのゾーンクラスタの一部として明示的に構成されるファイルシステム、ZFS データセット、ネットワークリソースなどの要素のみが含まれます。ゾーンクラスタ内のフ

ェイルオーバーアプリケーションは、ゾーンクラスタ内のあるノードから、同一ゾーンクラスタ内の別のノードにのみ、フェイルオーバーまたはスイッチオーバーすることができます。スケーラブルなアプリケーションのインスタンスはすべて、同一のゾーンクラスタ内でのみ実行されます。ゾーンクラスタは、アプリケーションがエスケープできないセキュリティーコンテナです。

- リソース管理 – Solaris のリソース管理制御の全機能を、ゾーンクラスタに対して適用できます。したがって、ゾーンクラスタ内のあるノード上に存在するすべてのアプリケーションを、ゾーンレベルで制御できます。これにより、ゾーンクラスタノードで利用できるリソースを、より効率的に管理できます。たとえば、1 つのゾーンクラスタ内に 1 つのアプリケーションを配置して、CPU の数を減らすことができます。こうすることで、CPU ごとのライセンス料金を削減できます。
- 委任管理 – ゾーンクラスタ内のアプリケーションを管理する権限を、そのゾーンクラスタで処理を行っている管理者に委任できます。ゾーンクラスタは、グローバルクラスタやその他のゾーンクラスタとは独立して機能します。大域ゾーンの管理者は、ゾーンクラスタ内のクラスタ間の依存関係やアフィニティーの設定、およびアプリケーションの管理を行うことができます。
- 単純化されたクラスタ – ゾーンクラスタ内で行う必要があるのは、アプリケーションおよびそのアプリケーションが使用するリソースの管理のみです。大域ゾーンの管理者は、ゾーンクラスタの内部および外部でコマンドを実行することにより、いつでもゾーンクラスタを作成、管理および削除できます。これらの操作は、グローバルクラスタやその他のゾーンクラスタには影響しません。

クラスタインターコネクト

クラスタインターコネクトは、クラスタ内の Solaris ホスト間でクラスタプライベート通信やデータサービス通信を転送する物理的なデバイス構成です。

冗長なインターコネクトの 1 つに障害が発生しても、操作は残りのインターコネクトを使って続けられます。そのため、システム管理者は、その間に障害を分離し、通信を修復することができます。Sun Cluster ソフトウェアは障害を検知し、修復し、修復されたインターコネクト経由の通信を自動的に再始動します。

詳細については、41 ページの「クラスタインターコネクトコンポーネント」を参照してください。

クラスタメンバーシップ

クラスタメンバーシップモニター (CMM) は、クラスタインターコネクトを使ってメッセージを交換し、次の処理を行う一連の分散エージェントです。

- すべてのノード (定足数) で一貫したメンバーシップの表示を行います。
- メンバーシップの変更に応じて同期のとれた再構成を行います。
- クラスタのパーティション分割を処理します。
- 障害のあるノードを、障害が修復されるまでクラスタから除外することによって、すべてのクラスタメンバー間の完全な接続を維持します。

CMM の主な機能はクラスタメンバーシップを確立することですが、そのためには、クラスタに逐次参加するノード群に関してクラスタ全体が合意していなければなりません。CMM は、1 つまたは複数のノード間での通信の途絶など、各ノードにおけるクラスタステータスの大きな変化を検知します。CMM は、トランスポートカーネルモジュールを使ってハートビートを生成し、トランスポート媒体を通してそれをクラスタのほかのノードに伝送します。定義されたタイムアウト時間内にノードからハートビートが送られてこないと、CMM は、そのノードに障害が発生したものとみなし、クラスタの再構成を通してクラスタメンバーシップの再設定を試みます。

CMM は、クラスタメンバーシップを確定し、データの整合性を確保するために、次の処理を行います。

- クラスタへのノードの参加、またはクラスタからのノードの脱退など、クラスタメンバーシップの変更を把握します。
- 異常のあるノードを、クラスタから切り離された状態に保ちます。
- 異常のあるノードを、それが修復されるまで非アクティブの状態に保ちます。
- クラスタそのものがノードのサブセットに分割されないように防止します。

クラスタ自身が複数の異なるクラスタに分割されないようにする方法についての詳細は、24 ページの「split-brain と amnesia」を参照してください。

クラスタ構成レポジトリ

クラスタ構成レポジトリ (Cluster Configuration Repository、CCR) は、クラスタの構成や状態に関する情報を格納するための、クラスタ全体に有効なプライベート分散データベースです。構成データを破損しないために、個々のノードは、クラスタリソースの現在の状態を知っている必要があります。この CCR のおかげで、すべてのノードが、一貫性のあるクラスタ像を持つことができます。CCR は、エラーや復旧の情況が発生したり、クラスタの一般的なステータスに変化があると更新されます。

CCR 構造には、次のような情報が含まれています。

- クラスタ名とノード名
- クラスタトランスポート構成
- Solaris Volume Manager ディスクセットや Veritas ディスクグループの名前
- 個々のディスクグループをマスターできるノードのリスト
- データサービスの操作に関するパラメータ値
- データサービスコールバックメソッドへのパス
- DID デバイス構成
- クラスタの現在のステータス

定足数デバイス

定足数デバイスとは、複数のノードによって共有される共有ストレージデバイスまたは定足数サーバーで、定足数を確立するために使用される票を構成します。クラスタは、票の定足数が満たされた場合にのみ動作可能です。定足数デバイスは、クラスタが独立したノードの集合にパーティション分割されたときに、どちらのノード集合が新しいクラスタを構成するかを確定するために使用されます。

クラスタノードと定足数デバイスはどちらも、定足数を確立するために投票します。デフォルトにより、クラスタノードは、起動してクラスタメンバーになると、定足数投票数 (quorum vote count) を 1 つ獲得します。ノードは、ノードのインストール中や管理者がノードを保守状態にした時には、投票数は 0 になります。

定足数デバイスは、デバイスへのノード接続の数に基づいて投票数を獲得します。定足数デバイスは、設定されると、最大投票数 *N*-1 を獲得します。この場合、*N* は、定足数デバイスへ接続された投票数を示します。たとえば、2 つのノードに接続された、投票数がゼロ以外の定足数デバイスの投票数は 1 (2–1) になります。

フォルトモニター

Sun Cluster システムでは、アプリケーションそのものや、ファイルシステム、ネットワークインタフェースを監視することによって、ユーザーとデータ間の「パス」にあるすべてのコンポーネントの高い可用性を保ちます。

Sun Cluster ソフトウェアは、ノードを素早く検知し、そのノードと同等のリソースを備えたサーバーを作成します。Sun Cluster ソフトウェアのおかげで、障害のあるノードの影響を受けないリソースはこの復旧中も引き続き使用され、障害のあるノードのリソースは復旧すると同時に再び使用可能になります。

データサービス監視

Sun Cluster の各データサービスには、データサービスを定期的に検査してその状態を判断するフォルトモニターがあります。フォルトモニターは、アプリケーションデーモンが動作しているかどうかや、クライアントにサービスが提供されているかどうかを検証します。探索によって得られた情報をもとに、デーモンの再起動やフェイルオーバーの実行などの事前に定義された処置が開始されます。

ディスクパスの監視

Sun Cluster ソフトウェアは、ディスクパス監視 (DPM) がサポートします。DPM は、二次ディスクパスの障害を報告することによって、フェイルオーバーやスイッチオーバーの全体的な信頼性を高めます。ディスクパスの監視には 2 つの方法があります。1 つめの方法は `cldevice` コマンドを使用する方法です。このコマンドを使用すると、クラスタ内のディスクパスの状態を監視、監視解除、または表示できます。コマンド行オプションについての詳細は、`cldevice(1CL)` のマニュアルページを参照してください。

2 つめの方法は、Sun Cluster Manager の GUI (Graphical User Interface) を使用してクラスタ内のディスクパスを監視する方法です。Sun Cluster Manager では、監視されているディスクパスがトポロジで表示されます。このトポロジビューは 10 分ごとに更新され、失敗した ping の数が表示されます。

IP マルチパス監視

クラスタの各 Solaris ホストは、そのクラスタのほかのホストの構成とは異なる、独自の IP ネットワークマルチパス 構成を持ちます。IP ネットワークマルチパス は、次のネットワークの通信障害を監視します。

- ネットワークアダプタの送信/受信パスがパケットの伝送を停止した。
- ネットワークアダプタとリンクとの接続がダウンしている。
- Ethernet スイッチ上のポートがパケットを送受信しない。
- グループ内の物理インタフェースがシステムの起動時に存在しない。

定足数デバイス監視

Sun Cluster ソフトウェアは、定足数デバイスの監視をサポートしています。クラスタ内の各ノードは周期的に、ローカルノードと構成されている各定足数デバイスとが正常に連携しているかどうかをテストします。構成されている定足数デバイスとは、そのローカルノードに対する構成パスを持ち、かつ保守モードでないデバイスをいいます。このテストでは、定足数デバイスの定足数キーの読み込みを試みます。

Sun Cluster システムは、以前は正常だった定足数デバイスが障害を起こしているのを発見すると、自動的にその定足数デバイスを異常としてマーク付けします。以前は異常だった定足数デバイスが正常に戻っているのを発見すると、自動的にその定足数デバイスを正常としてマーク付けし、その定足数デバイスに適切な定足数情報を配置します。

Sun Cluster システムは、定足数デバイスが正常かどうかのステータスが変更された場合にレポートを生成します。ノードを再構成するとき、異常な定足数デバイスはメンバーシップの票を構成できません。したがって、そのクラスタは処理を継続できない可能性があります。

データの完全性

Sun Cluster システムはデータ破損を防ぎ、データの完全性を保とうとします。それぞれのクラスタノードはデータとリソースを共有していますので、クラスタが、同時にアクティブである複数のパーティションに分割されることがあってはなりません。CMM は、必ず 1 つのクラスタだけが使用可能であることを保証します。

split-brain と amnesia

クラスタのパーティション分割によって起こる問題に、*split-brain* と *amnesia* の 2 つがあります。split-brain が起こるのは、Solaris ホスト間のクラスタインターコネクトが失われてクラスタがサブクラスタにパーティション分割され、それぞれのサブクラスタがそれを唯一のパーティションであると認識する場合です。ほかのサブクラスタの存在を認識していないサブクラスタは、ネットワークアドレスの重複やデータ破損など、共有リソースの対立を引き起こすおそれがあります。

amnesia は、すべてのノードがそのクラスタ内で不安定なグループの状態になっている場合に起こります。たとえば、ノード A とノード B からなる 2 ノードクラスタがあるとします。ノード A が停止すると、CCR の構成データはノード B のものだけが更新され、ノード A のものは更新されません。この後でノード B が停止し、ノード A が再起動されると、ノード A は CCR の古い内容に基づいて動作することになります。この状態を amnesia と呼びます。この状態になると、クラスタは、古い構成情報で実行されることがあります。

split-brain と amnesia の問題は、各ノードに 1 票を与え、過半数の投票がないとクラスタが動作しないようにすることで防止できます。過半数の投票を得たパーティションは「定足数 (quorum)」を獲得し、アクティブになります。この過半数の投票メカニズムは、クラスタのノード数が 2 を超える場合には有効です。しかし、2 ノードクラスタでは過半数が 2 であるため、このようなクラスタがパーティション分割されると、パーティションは外部からの投票で定足数を獲得します。この外部からの投票は、定足数デバイスによって行われます。定足数デバイスは、2 つのノードで共有されている任意のディスクにすることができます。

表 2–1に、Sun Cluster ソフトウェアが定足数を使用して split-brain と anmesia を回避する様子を示します。

表 2–1　クラスタ定足数、および split-brain と amnesia の問題

問題	定足数による解決策
split brain	過半数の投票を獲得したパーティション (サブクラスタ) だけをクラスタとして実行できるようにする (過半数を獲得できるパーティションは 1 つのみ)。ノードが定足数を獲得できないと、ノードはパニックになります。
amnesia	起動されたクラスタには、最新のクラスタメンバーシップのメンバーであった (したがって、最新の構成データを持つ) ノードが少なくとも 1 つあることを保証する。

フェンシング

split brain が発生すると、一部のノードが通信できなくなるため、個々のノードまたは一部のノードが個々のクラスタまたは一部のクラスタを形成しようとします。各部分、つまりパーティションは、多重ホストディスクに対して単独のアクセスと所有権を持つものと誤って認識します。しかし、複数のノードがディスクに書き込もうとすると、データ破損を招くおそれがあります。

ノードは、ほかのノードとの接続を失うと、通信が可能なノードとクラスタを形成しようとします。そうしたノードの集合が定足数を得られない場合、Sun Cluster ソフトウェアはそのノードを停止して、ディスクから「フェンシング」します。つまり、そのノードがディスクにアクセスできないようにします。現在のメンバーノードだけが、ディスクへのアクセス権を持つため、データの完全性が保たれます。

フェンシングは、選択したディスクまたはすべてのディスクに対して無効にすることができます。

注意 – 不適切な状況でフェンシングを無効にすると、アプリケーションのフェイルオーバー時にデータが破損する危険性が高くなります。フェンシングの無効化を検討する場合には、データ破損の可能性を十分に調査してください。共有ストレージデバイスで SCSI プロトコルがサポートされていない場合 (Serial Advanced Technology Attachment (SATA) ディスクなど)、またはクラスタのストレージにクラスタ外部にあるホストからのアクセスを許可する場合に、フェンシングを無効にします。

フェイルファースト

フェイルファーストの目的は、正常な処理を継続できない、正常でないコンポーネントを停止することです。Sun Cluster ソフトウェアは、さまざまな異常な状態を検出するための、多くのフェイルファーストメカニズムを備えています。

Sun Cluster システムは、グローバルクラスタ投票ノードでクリティカルな障害を検出すると、強制的にその Solaris ホストをシャットダウンします。

その他の種類のノード (グローバルクラスタ非投票ノード、ゾーンクラスタノードなど) にクリティカルな障害を検出した場合は、そのノードを再起動します。

Sun Cluster ソフトウェアは、クラスタに属するノードを監視します。通信やノードの障害により、クラスタのノード数は変わります。クラスタが十分な投票数を維持できない場合、Sun Cluster ソフトウェアはそのノードの集合を停止します。

Sun Cluster ソフトウェアは、多数のクリティカルなクラスタ固有デーモンを維持管理します。デーモンには、グローバルクラスタ投票ノードをサポートするものや、その他の種類のノードをサポートするものがあります。デーモンは、そのデーモンがサポートするノードにとってクリティカルです。サポートするノードは、デーモンが実行されているノードによって異なります。たとえば、大域ゾーンの一部のデーモンは、非大域ゾーンをサポートします。このため、これらのデーモンは、大域ゾーンよりもむしろ非大域ゾーンにとってクリティカルになります。

グローバルデバイス、ローカルデバイス、およびデバイスグループ

クラスタファイルシステムでは、クラスタのすべてのファイルがすべてのノードから同じように認識され、アクセス可能になります。それと同様に、Sun Cluster ソフトウェアの下では、クラスタのすべてのデバイスがクラスタ全体から認識され、アクセス可能になります。つまり、どのノードからでも入出力サブシステムを通してクラスタのどのデバイスにもアクセスできます。デバイスが物理的にどこに接続されているかは関係ありません。このアクセスをグローバルデバイスアクセスと呼びます。

グローバルデバイス

Sun Cluster システムでは、クラスタの任意のデバイスに任意のノードから高い可用性をもってクラスタレベルでアクセスできるようにするために、グローバルデバイスを使用します。

Sun Cluster でグローバルデバイスを使用する方法

通常、ノードからグローバルデバイスにアクセスできないことがあると、Sun Cluster ソフトウェアは、そのデバイスへのパスを別のパスに切り替え、アクセスをそのパスに振り向けます。グローバルデバイスでは、この変更は簡単です。どのパスを使用する場合でも、デバイスには同じ名前が使用されるからです。リモートデバイスへのアクセスは、同じ名前を持つローカルデバイスの場合と同じように行われます。さらに、クラスタのグローバルデバイスにアクセスする API は、ローカルのデバイスにアクセスする API と同じです。

Sun Cluster グローバルデバイスには、ディスク、CD-ROM、テープが含まれます。ただし、サポートされるマルチポートのグローバルデバイスはディスクだけです。つまり、CD-ROM とテープは、現在可用性の高いデバイスではありません。各サーバーのローカルディスクも多重ポート化されていないため、可用性の高いデバイスではありません。

クラスタは、クラスタ内の各ディスク、CD-ROM、テープデバイスに一意の ID を割り当てます。この割り当てによって、クラスタ内の任意のノードから各デバイスに対して一貫したアクセスが可能になります。

デバイス ID

Sun Cluster ソフトウェアは、デバイス ID (DID) ドライバと呼ばれるコンストラクトを通してグローバルデバイスを管理します。このドライバを使用して、多重ホストディスク、テープドライブ、CD-ROM を含め、クラスタ内のあらゆるデバイスに一意の ID を自動的に割り当てます。

DID ドライバは、クラスタのグローバルデバイスアクセス機能の重要な部分です。DID ドライバは、クラスタのすべてのノードを検査し、一意のディスクデバイスからなるリストを構築します。さらに、DID ドライバは、一意のメジャー番号とマイナー番号を各デバイスに割り当てます。この数字は、クラスタのすべてのノードで一貫性をもって管理されます。グローバルデバイスへのアクセスは、従来の Solaris DID と替わって DID ドライバによって割り当てられた一意の DID を使って行われます。

このような方法をとることで、Solaris Volume Manager など、ディスクにアクセスするアプリケーションが何であれ、クラスタ全体で一貫性のあるパスが使用されます。多重ホストディスクの場合は、この一貫性がとりわけ重要です。各デバイスのローカルのメジャー番号とマイナー番号はノードによって異なる可能性があるからです。さらに、これらの数字は、Solaris デバイスの命名規約も同様に変更する可能性があります。

ローカルデバイス

Sun Cluster ソフトウェアはローカルデバイスも管理します。このようなデバイスは、サービスを実行していてクラスタに物理的に接続されている Solaris ホストでのみアクセス可能です。ローカルデバイスは、性能の点でグローバルデバイスよりも有利です。ローカルデバイスでは、状態情報を複数のホストに同時に複製する必要がないからです。デバイスのドメインに障害が発生すると、そのデバイスにはアクセスできなくなります。ただし、そのデバイスを複数のホストで共有できる場合を除きます。

デバイスグループ

デバイスグループは、ボリュームマネージャーのディスクグループを「グローバル」デバイスにします。デバイスグループは、使用しているディスクに対してマルチパスと多重ホストをサポートするからです。多重ホストディスクに物理的に接続された各クラスタの Solaris ホストは、デバイスグループへのパスを提供します。

Sun Cluster システムで Sun Cluster ソフトウェアを使用している多重ホストディスクを制御するには、多重ホストディスクをデバイスグループとして登録します。この登録によって、Sun Cluster システムは、どのノードがどのボリュームマネージャーディスクグループへのパスを持っているかを知ることができます。Sun Cluster ソフトウェアは、クラスタ内のディスクデバイスやテープデバイスごとに、raw デバイスグループを作成します。これらのクラスタデバイスグループは、ユーザーがクラスタファイルシステムをマウントするか、raw データベースファイルにアクセスすることによって、これらのデバイスグループにグローバルデバイスとしてアクセスするまで、オフライン状態に置かれます。

データサービス

データサービスは、Sun Cluster 構成の下でアプリケーションを変更なしで実行できるようにする、ソフトウェアと構成ファイルの組み合わせです。Sun Cluster 構成の下で動作するアプリケーションは、リソースグループマネージャー (Resource Group Manager、RGM) の制御下にある 1 つのリソースです。データサービスを使えば、Sun Java System Web Server や Oracle データベースなどのアプリケーションをクラスタで (単一のサーバーではなく) 実行するように構成できます。

データサービスのソフトウェアは、アプリケーションに対して次の操作を行う Sun Cluster 管理メソッドを実装しています。

- アプリケーションの起動
- アプリケーションの停止
- アプリケーションの障害の監視とこの障害からの復旧

データサービスの構成ファイルは、RGM にとってアプリケーションを意味するリソースのプロパティーを定義したものです。

クラスタのフェイルオーバーデータサービスやスケーラブルデータサービスの処理はRGMによって制御されます。RGMは、クラスタメンバーシップの変更に応じて選択されたクラスタのノードでデータサービスの起動や停止を行います。データサービスアプリケーションは、RGMを通してクラスタフレームワークを利用できます。

RGMはデータサービスをリソースとして制御します。これらの実装はSunによって提供されるか、開発者によって作成されます。後者の場合には、汎用的なデータサービステンプレートや、データサービス開発ライブラリAPI (Data Service Development Library API、DSDL API)、リソース管理API (Resource Management API、RMAPI)が使用されます。クラスタ管理者は、リソースグループと呼ばれる入れ物(コンテナ)の中でリソースの作成や管理を行います。リソースやリソースグループの状態は、RGMや管理者のアクションによってオンラインやオフラインにされます。

リソースタイプの説明

リソースタイプとは、あるアプリケーションをクラスタに説明するプロパティーの集まりのことです。この集合には、クラスタのノードでアプリケーションをどのように起動、停止、監視するかを示す情報が含まれています。さらに、リソースタイプには、アプリケーションをクラスタで使用するために必要なアプリケーション固有のプロパティーも含まれています。Sun Clusterデータサービスには、いくつかのリソースタイプが事前に定義されています。たとえば、Sun Cluster HA for Oracleのリソースタイプは`SUNW.oracle-server`、Sun Cluster HA for Apacheのリソースタイプ`SUNW.apache`です。

リソースの説明

リソースとは、クラスタ規模で定義したリソースタイプのインスタンスのことです。リソースタイプを使用すると、アプリケーションの複数のインスタンスをクラスタにインストールできます。ユーザーがリソースを初期化すると、RGMは、アプリケーション固有のプロパティーに値を割り当てます。リソースは、リソースタイプのレベルにあるすべてのプロパティーを継承します。

データサービスは、いくつかのタイプのリソースを使用します。たとえば、Apache Web ServerやSun Java System Web Serverなどのアプリケーションは、それらが依存するネットワークアドレス(論理ホスト名と共有アドレス)を使用します。アプリケーションとネットワークリソースはRGMが管理する基本単位です。

リソースグループの説明

RGM は、複数のリソースをリソースグループという 1 つの単位として扱うことができるようにします。リソースグループとは、関連する (あるいは、相互に依存する) リソースの集合のことです。たとえば、SUNW.LogicalHostname リソースタイプから派生したリソースは、Oracle データベースリソースタイプから派生したリソースと同じリソースグループに置かれることがあります。リソースグループ上でフェイルオーバーまたはスイッチオーバーが開始されると、リソースグループは 1 つの単位として移行されます。

データサービスのタイプ

データサービスを使用すると、アプリケーションは可用性の高いものやスケーラブルなサービスになります。クラスタで単一の障害が発生した場合、大幅なアプリケーションの中断を回避できます。

データサービスを構成する際には、次のデータサービスのタイプから 1 つを選択する必要があります。

- フェイルオーバーデータサービス
- スケーラブルデータサービス
- パラレルデータサービス

フェイルオーバーデータサービスの説明

フェイルオーバーとは、クラスタがアプリケーションを障害のある稼動系から、指定の冗長化された待機系に自動的に再配置するプロセスのことをいいます。フェイルオーバーアプリケーションには、次の特徴があります。

- クラスタの 1 つのノードだけに実行の資格があります。
- クラスタで動作していることを意識させません。
- クラスタフレームワークに基づいて HA を達成します。

フォルトモニターは、エラーを検出すると、データサービスの構成に従って、同じノードでそのインスタンスを再起動しようとするか、別のノードでそのインスタンスを起動 (フェイルオーバー) しようとします。フェイルオーバーサービスは、アプリケーションインスタンスリソースとネットワークリソース (論理ホスト名) のコンテナである、フェイルオーバーリソースグループを使用します。論理ホスト名とは、1 つのノードに構成して、後で自動的に元のノードや別のノードに構成できる IP アドレスのことです。

サービスが一時的に中断されるため、クライアントは、フェイルオーバーの完了後にサービスに再接続しなければならない場合があります。しかし、クライアントは、サービスの提供元である物理サーバーが変更したことを意識しません。

スケーラブルデータサービスの説明

スケーラブルデータサービスでは、複数のアプリケーションインスタンスが複数のノードで同時に動作します。スケーラブルサービスは、2つのリソースグループを使用します。スケーラブルリソースグループにはアプリケーションリソースが、フェイルオーバーリソースグループには、スケーラブルサービスが依存するネットワークリソース (共有アドレス) がそれぞれ含まれています。スケーラブルリソースグループは、複数のノードでオンラインにできるため、サービスの複数のインスタンスを同時に実行できます。共有アドレスのホストとなるフェイルオーバーリソースグループは、一度に1つのノードでしかオンラインにできません。スケーラブルサービスをホストとするすべてのノードは、サービスをホストするための同じ共有アドレスを使用します。

クラスタは、同一のネットワークインタフェース (グローバルインタフェース) を通してサービス要求を受け取ります。これらの要求は、事前に定義されたいくつかのアルゴリズムの1つに基づいてノードに分配されます (アルゴリズムは負荷均衡ポリシーによって設定される)。クラスタは、負荷均衡ポリシーを使用し、いくつかのノード間でサービス負荷均衡をとることができます。

並列アプリケーションの説明

Sun Cluster システムは、パラレルデータベースを使用することによってクラスタのすべてのノードでアプリケーションを並列で実行できるようにする環境を提供します。Sun Cluster Support for Oracle Real Application Clusters は、インストールされている場合、Oracle Real Application Clusters を Sun Cluster ノードで実行できるようにするパッケージ群です。さらに、このデータサービスでは、Sun Cluster コマンドを使って Sun Cluster Support for Oracle Real Application Clusters を管理できます。

パラレルアプリケーションはクラスタ環境で動作するように考えられたものです。したがって、このようなアプリケーションは、複数のノードから同時マスターされます。Oracle Real Application Clusters 環境では、複数の Oracle インスタンスが協力して同じ共有データベースへのアクセス権を提供します。Oracle クライアントは、任意のインスタンスを使用してデータベースにアクセスできます。したがって、1つまたは複数のインスタンスで障害が発生しても、クライアントは残りのインスタンスに接続することによって、引き続きデータベースにアクセスできます。

システムリソースの使用状況

システムリソースは、CPU 使用率、メモリーの使用量、スワップの使用量、およびディスクとネットワークのスループットに関係します。

Sun Cluster ソフトウェアを使用すると、ノード、ディスク、ネットワークインタフェース、Sun Cluster のリソースグループ、Solaris ゾーンなどのオブジェクトタイプにより特定のシステムリソースがどれだけ使用されているかを監視できます。システムリソースの使用状況の監視は、リソース管理ポリシーの一部とすることができま

す。また、Sun Cluster では、あるリソースグループに割り当てられた CPU を制御し、リソースグループが実行されるプロセッサセットのサイズを制御できます。

システムリソース監視

Sun Cluster ソフトウェアを通してシステムリソースの使用状況を監視することにより、特定のシステムリソースを使用するサービスがどのように実行されているかを反映するデータを収集したり、リソースのボトルネックや過負荷などを見つけたりすることができ、これによって問題に事前に対処し、ワークロードをより効率的に管理することができます。システムリソースの使用状況に関するデータは、どのハードウェアリソースが十分に利用されておらず、どのアプリケーションが大量のリソースを使用しているかを判別するのに役立ちます。このデータに基づき、必要なリソースを備えたノードにアプリケーションを割り当てたり、フェイルオーバー先にするノードを選択したりできます。この統合により、ハードウェアリソースとソフトウェアリソースの使用方法を最適化できます。

あるデータの値がシステムリソースにとってクリティカルであると考えられる場合に、この値のしきい値を設定できます。しきい値を設定する際には、しきい値に重要度を割り当てることにより、このしきい値のクリティカル度の選択も行います。しきい値を超えると、Sun Cluster はそのしきい値の重要度を、ユーザーが選択した重要度に変更します。データ収集およびしきい値の構成についての詳細は、『Sun Cluster のシステム管理 (Solaris OS 版)』の第 9 章「CPU 使用率の制御の構成」を参照してください。

CPU の制御

クラスタ上で動作する各アプリケーションおよびサービスには、それぞれ固有の CPU の要件があります。表 2–2 に、Solaris オペレーティングシステムのさまざまなバージョンで利用可能な CPU 制御操作を示します。

表 2–2　CPU の制御

Solaris のバージョン	ゾーン	制御
Solaris 9 オペレーティングシステム	該当なし	CPU の配分の割り当て
Solaris 10 オペレーティングシステム	大域	CPU の配分の割り当て

表 2–2　CPU の制御　　(続き)

Solaris のバージョン	ゾーン	制御
Solaris 10 オペレーティングシステム	非大域	CPU の配分の割り当て CPU の数の割り当て 専用のプロセッサセットの作成

注 – CPU の配分を行う場合、クラスタ上のデフォルトのスケジューラを Fair Share Scheduler (FSS) にする必要があります。

非大域ゾーンで専用のプロセッサセットのリソースグループに割り当てられた CPU を制御すると、CPU をもっとも厳密に制御できます。これは、あるリソースグループ向けに CPU を予約した場合、この CPU はほかのリソースグループからは使用できないためです。CPU 制御の構成については、『Sun Cluster のシステム管理 (Solaris OS 版)』の第 9 章「CPU 使用率の制御の構成」を参照してください。

システムリソースの使用状況の視覚化

システムリソースデータと CPU の帰属を視覚化する方法には、コマンド行を使用する方法と Sun Cluster Manager グラフィカルユーザーインタフェースを使用する方法の 2 つがあります。コマンドからの出力は、ユーザーが要求する監視データの表形式の表現になります。Sun Cluster Manager を使用すると、データをグラフィック形式で視覚化できます。監視することを選択したシステムリソースは、ユーザーが視覚化できるデータを決定します。

◆ ◆ ◆ 第 3 章

3 Sun Clusterのアーキテクチャー

Sun Cluster アーキテクチャーでは、一連のシステムが単一の大規模システムとして配備され、管理され、認識されます。

この章で説明する内容は次のとおりです。

Sun Clusterのハードウェア環境

クラスタは、次のハードウェアコンポーネントから構成されます。

- ローカルディスク (非共有) に接続された Solaris ホスト。クラスタの主要なコンピューティングプラットフォームです。
- 多重ホストストレージ。Solaris ホスト間で共有されるディスクです。
- テープや CD-ROM などのリムーバブルメディア。グローバルデバイスとして構成されます。
- クラスタインターコネクト。ノード間の通信チャネルとして使用されます。
- パブリックネットワークインタフェース。クライアントシステムによって使用されるネットワークインタフェースは、このインタフェースを通してクラスタのデータサービスにアクセスします。

図 3–1に、ハードウェアコンポーネント相互間の連携のしくみを示します。

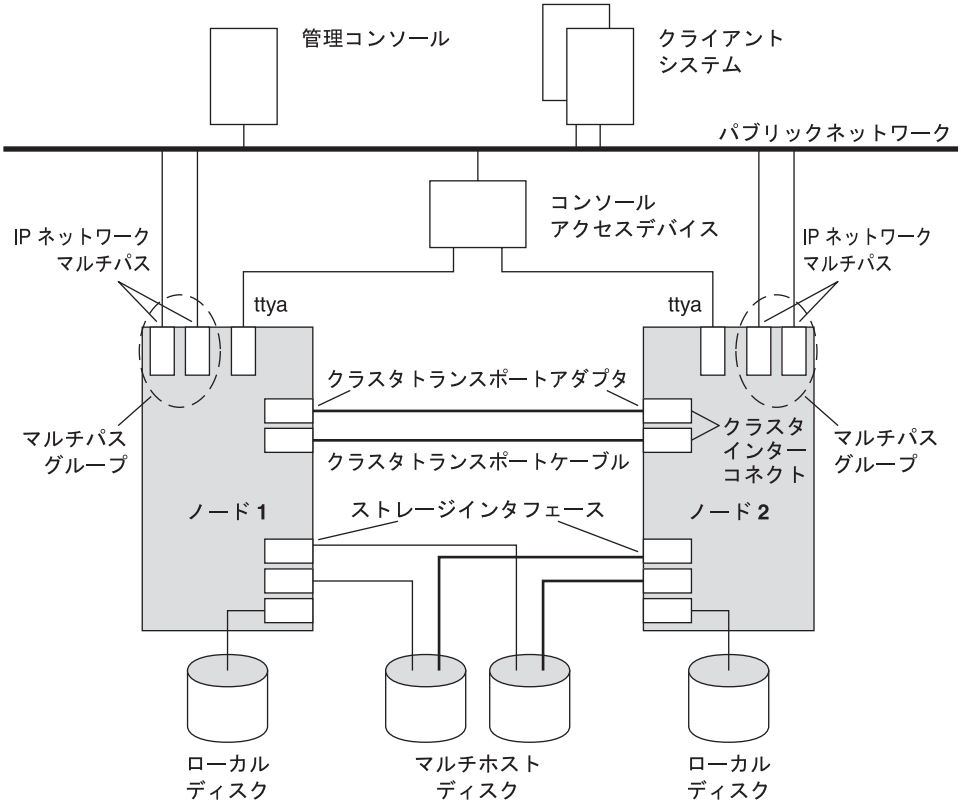

図 3–1　Sun Cluster ハードウェアコンポーネント

Sun Clusterのソフトウェア環境

Solaris ホストがクラスタメンバーとして動作するためには、ホストに次のソフトウェアがインストールされていなければなりません。

- Solaris ソフトウェア
- Sun Cluster ソフトウェア
- データサービスアプリケーション
- ボリューム管理 (Solaris™ Volume Manager または Veritas Volume Manager)

  ただし、そのボックス自体のボリューム管理を使用する構成は例外です。この構成では、ソフトウェアボリュームマネージャーが必要ない場合があります。

図 3–2に、相互に機能して Sun Cluster ソフトウェア環境を構成するソフトウェアコンポーネントの概要を示します。

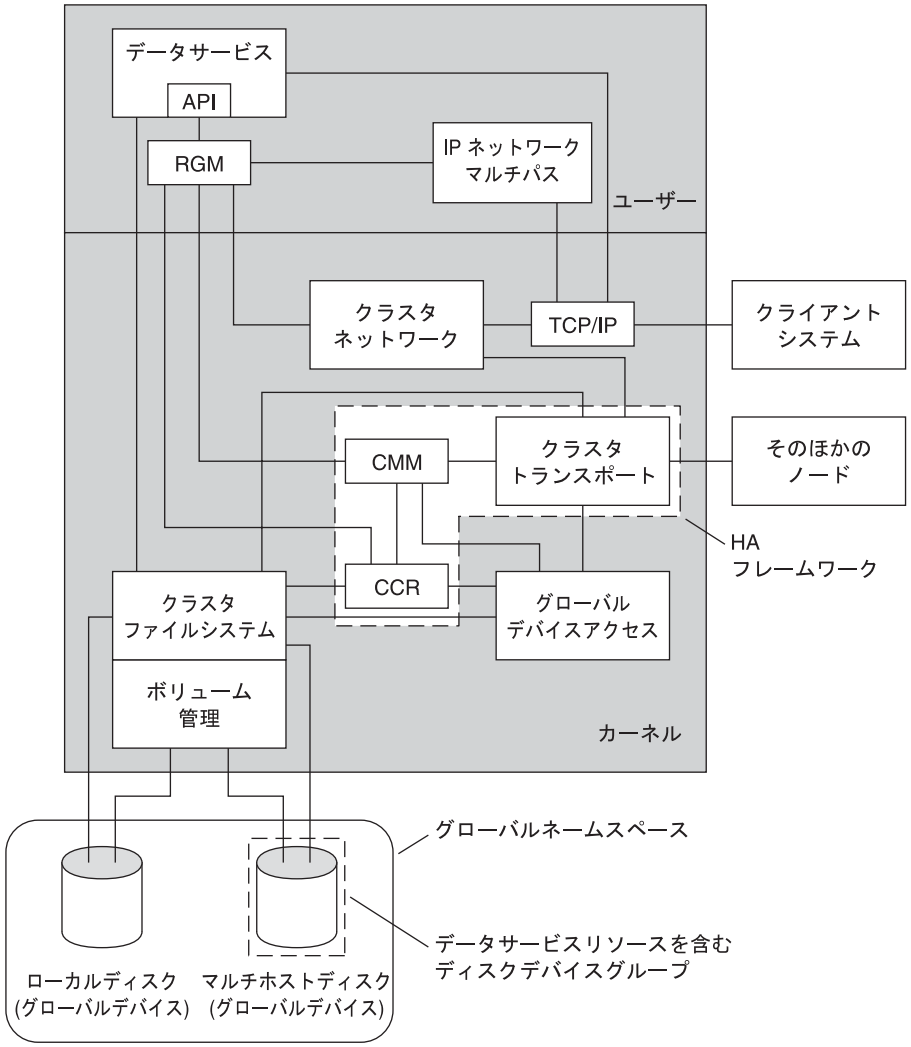

図 3–2　Sun Cluster ソフトウェアアーキテクチャー

クラスタメンバーシップモニター

データが破壊から保護されるように保証するには、すべてのノードが、クラスタメンバーシップに対して一定の同意に達していなければなりません。必要であれば、CMM は、障害に応じてクラスタサービスのクラスタ再構成を調整します。

CMM は、クラスタのトランスポート層から、他のノードへの接続に関する情報を受け取ります。CMM は、クラスタインターコネクトを使用して、再構成中に状態情報を交換します。

CMMは、クラスタメンバーシップの変更を検出すると、それに合わせてクラスタを構成します。この構成処理では、クラスタリソースが、クラスタの新しいメンバーシップに基づいて再配布されることがあります。

CMMは完全にカーネル内で動作します。

クラスタ構成レポジトリ (Cluster Configuration Repository、CCR)

CCRは、CMMに依存して、定足数 (quorum) が確立された場合にのみクラスタが実行されるように保証します。CCRは、クラスタ全体のデータの一貫性を確認し、必要に応じて回復を実行し、データへの更新を容易にします。

クラスタファイルシステム

クラスタファイルシステムは、次のコンポーネント間のプロクシです。

- あるSolarisホスト上のカーネルとそのホストが使用しているファイルシステム
- そのディスク (1つまたは複数) と物理的に接続されているSolarisホスト上のボリュームマネージャー

クラスタファイルシステムでは、グローバルデバイス (ディスク、テープ、CD-ROM) が使用されます。グローバルデバイスには、クラスタのどのSolarisホストからでも同じファイル名 (たとえば、`/dev/global/`) を使ってアクセスできます。そのホストは、アクセスするストレージデバイスに物理的に接続されている必要はありません。ユーザーは、グローバルデバイスを通常のデバイスと同じように使用できます。つまり、`newfs` や `mkfs` を使ってグローバルデバイスにファイルシステムを作成することができます。

クラスタファイルシステムには、次の機能があります。

- ファイルのアクセス場所が透過的になります。システムのどこにあるファイルでも、プロセスから開くことができます。さらに、すべてのホストのプロセスから同じパス名を使ってファイルにアクセスできます。

  注 – クラスタファイルシステムは、ファイルを読み取る際に、ファイル上のアクセス時刻を更新しません。

- 一貫したプロトコルを使用して、ファイルが複数のホストから同時にアクセスされる場合でも、UNIXファイルアクセスセマンティクスを維持します。
- 拡張キャッシュ機能とゼロコピーバルク入出力移動機能により、ファイルデータを効率的に移動することができます。

- クラスタファイルシステムには、fcntl(2) インタフェースに基づく、高度な可用性を備えたアドバイザリファイルロッキング機能があります。クラスタファイルシステムのファイルに対してアドバイザリファイルロッキング機能を使用することにより、複数のクラスタホストで動作するアプリケーションの間で、データのアクセスを同期化できます。ファイルロックを所有するノードがクラスタから切り離されたり、ファイルロックを所有するアプリケーションが異常停止すると、それらのロックはただちに解放されます。
- 障害が発生した場合でも、データへの連続したアクセスが可能です。アプリケーションは、ディスクへのパスが有効であれば、障害による影響を受けません。この保証は、raw ディスクアクセスとすべてのファイルシステム操作で維持されます。
- クラスタファイルシステムは、基本のファイルシステムからもボリュームマネージャーからも独立しています。クラスタシステムファイルは、サポートされているディスク上のファイルシステムすべてを広域にします。

スケーラブルデータサービス

クラスタネットワーキングの主な目的は、データサービスにスケーラビリティを提供することにあります。スケーラビリティーとは、サービスに提供される負荷が増えたときに、新しいノードがクラスタに追加されて新しいサーバーインスタンスが実行されるために、データサービスがこの増加した負荷に対して一定の応答時間を維持できるということを示します。スケーラブルデータサービスの例としては、Web サービスがあります。通常、スケーラブルデータサービスはいくつかのインスタンスからなり、それぞれがクラスタの異なるノードで実行されます。これらのインスタンスは、遠隔クライアントに対して単一のサービスとして動作し、そのサービスの機能を提供します。別々のノードで動作するいくつかの httpd デーモンからなるスケーラブル Web サービスでは、任意のデーモンでクラスタ要求を処理できます。要求に対応するデーモンは、負荷分散ポリシーによって決められます。クライアントへの応答は、その要求にサービスを提供する特定のデーモンからではなく、サービスからのもののようにみえるため、単一サービスの外観が維持されます。

次の図は、スケーラブルサービスの構造を示したものです。

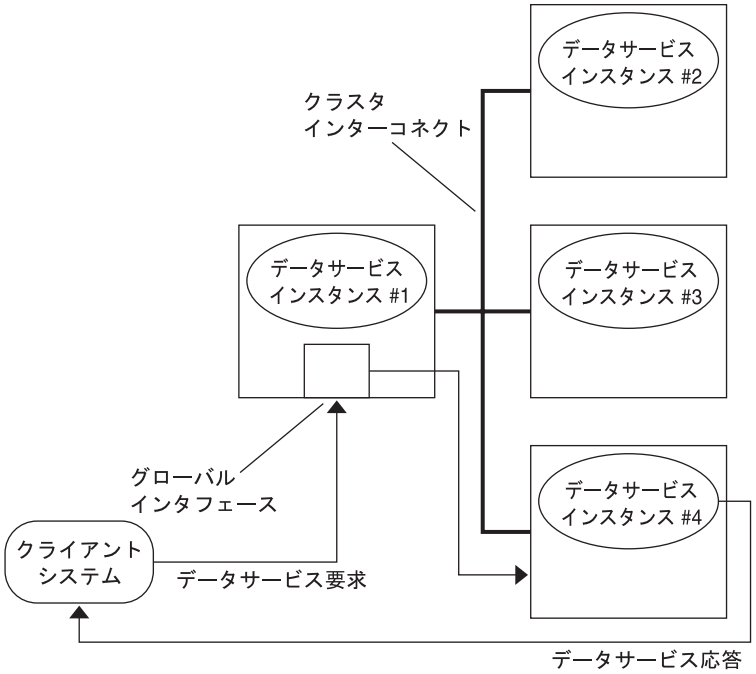

図3–3 スケーラブルデータサービスの構造

グローバルインタフェースのホストではないノード (プロキシノード) には、そのループバックインタフェースでホストされる共有アドレスがあります。グローバルインタフェースに入ってくるパケットは、構成可能な負荷均衡ポリシーに基づいてほかのクラスタノードに分配されます。次に、構成できる負荷均衡ポリシーについて説明します。

負荷均衡ポリシー

負荷均衡は、スケーラブルサービスのパフォーマンスを応答時間とスループットの両方の点で向上させます。

スケーラブルデータサービスには、*pure* と *sticky* の 2 つのクラスがあります。pure サービスとは、どのインスタンスでもクライアント要求に応答できるサービスをいいます。sticky サービスでは、ノードへの要求の負荷をクラスタが均衡させます。これらの要求は、別のインスタンスには変更されません。

pure サービスは、ウェイト設定した (weighted) 負荷均衡ポリシーを使用します。この負荷均衡ポリシーのもとでは、クライアント要求は、デフォルトで、クラスタ内のサーバーインスタンスに一律に分配されます。たとえば、各ノードのウェイトが 1 であるような 3 ノードクラスタでは、各ノードが、任意のクライアントからの要求

をそのサービスのために 3 分の 1 ずつ処理します。ウェイトの変更は、`clresource(1cl)` コマンドインタフェースか Sun Cluster Manager GUI を使っていつでもできます。

sticky サービスには、*ordinary sticky* と *wildcard sticky* があります。sticky サービスを使用すると、内部状態メモリーを共有でき (アプリケーションセッション状態)、複数の TCP 接続でアプリケーションレベルの同時セッションが可能です。

ordinary sticky サービスを使用すると、クライアントは、複数の同時 TCP 接続で状態を共有できます。このクライアントを、単一ポートで待機するサーバーインスタンスに対して「sticky」であるといいます。クライアントは、インスタンスが起動していてアクセス可能であり、負荷均衡ポリシーがサービスのオンライン時に変更されていなければ、すべての要求が同じサーバーのインスタンスに送られることを保証されます。

wildcard sticky サービスは、動的に割り当てられたポート番号を使用しますが、クライアント要求が同じノードに送りかえされると想定します。クライアントは、同じ IP アドレスに対して、複数のポート間で sticky wildcard であるといいます。

多重ホストディスク記憶装置

Sun Cluster ソフトウェアは、複数のノードに同時に接続できる多重ホストディスクストレージを使用することによって、ディスクの高い可用性を実現します。これらのディスクは、ボリューム管理ソフトウェアの使用を通して、クラスタノードからマスターされる共有ストレージに編成されます。そして、障害が発生したときに別のノードに移動されるように構成されます。Sun Cluster システムで多重ホストディスクを使用することには、さまざまな利点があります。たとえば、次はその例です。

- ファイルシステムへのグローバルアクセス
- ファイルシステムやデータへの複数のアクセスパス
- 単一ノード障害に耐えられる

クラスタインターコネクトコンポーネント

1 つのクラスタには、1 - 6 つまでのクラスタインターコネクトを設定できます。クラスタインターコネクトを 1 つだけ使用すると、プライベートインターコネクトに使用されるアダプタポートの数が減りますが、同時に冗長性がなくなり、可用性が低くなります。また、その 1 つのインターコネクトに障害が発生すると、クラスタが自動回復を実行するのによけいに時間がかかります。クラスタインターコネクトが 2 つ以上になると冗長性とスケーラビリティーが提供されるので、シングルポイント障害が回避されて可用性も高くなります。

Sun Cluster インターコネクトでは、Fast Ethernet、Gigabit-Ethernet、InfiniBand、または Scalable Coherent Interface (SCI, IEEE 1596-1992) の使用を通して、高性能のクラスタプライベート通信がサポートされます。

クラスタ環境のノード間通信には、高速、低遅延のインターコネクトとプロトコルが欠かせません。Sun Cluster システムの SCI インターコネクトは、一般的なネットワークインタフェースカード (NIC) よりも高い性能を発揮します。

RSM Reliable Datagram Transport (RSMRDT) ドライバは、RSM API 上に構築されるドライバと、RSMRDT-API インタフェースをエクスポートするライブラリから構成されます。このドライバは、Oracle Real Application Clusters の性能を向上させます。このドライバはまた、負荷均衡機能と高可用性 (High-Availability、HA) 機能をドライバ内部で直接提供することにより、両機能を強化すると共に、クライアントからの透過な利用を可能にしています。

クラスタインターコネクトは、以下のハードウェアコンポーネントで構成されます。

- アダプタ – 個々のクラスタホストに存在するネットワークインタフェースカード。複数のインタフェースを持つネットワークアダプタは、アダプタ全体に障害が生じると、単一地点による障害の原因となる可能性があります。
- スイッチ – ジャンクションとも呼ばれる、クラスタホストの外部にあるスイッチ。スイッチは、パススルーおよび切り換え機能を実行して、3 つ以上のホストへの接続を可能にします。2 ホストクラスタでは、冗長な物理ケーブルによってホストを相互に直接接続できるため、アダプタハードウェアでスイッチが必要な場合を除き、スイッチは必要ありません。これらの冗長なケーブルは、各ホストの冗長化されたアダプタに接続されます。ホストを 3 つ以上使用する構成では、スイッチが必要です。
- ケーブル – 2 つのネットワークアダプタまたはアダプタとスイッチの間をつなぐ物理接続。

図 3–4 に、3 つのコンポーネントがどのように接続されているかを示します。

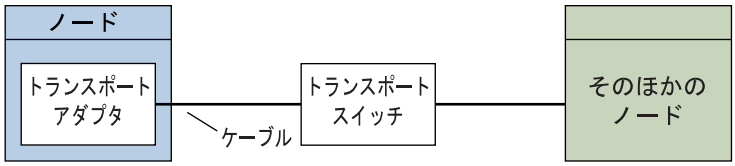

図 3–4 クラスタインターコネクト

IP ネットワークマルチパスグループ

パブリックネットワークアダプタは、IPMP グループ (マルチパスグループ) として編成されます。各マルチパスグループには、1 つまたは複数のパブリックネットワークアダプタがあります。マルチパスグループの各アダプタはアクティブにすることができます。あるいは、スタンバイインタフェースを構成し、フェイルオーバーが起こるまでそれらを非アクティブにしておくことができます。

マルチパスグループは、論理ホスト名と共有アドレスリソースの基盤です。つまり、ノード上の同じマルチパスグループは、任意の数の論理ホスト名または共有アドレスリソースをホストできます。マルチパスを作成すれば、クラスタノードのパブリックネットワーク接続を監視できます。

論理ホスト名や共有アドレスリソースについては、『Sun Cluster データサービスの計画と管理 (Solaris OS 版)』を参照してください。

パブリックネットワークインタフェース

クライアントは、パブリックネットワークインタフェースを介してクラスタに接続します。各ネットワークアダプタカードは、カードに複数のハードウェアインタフェースがあるかどうかによって、1 つまたは複数のパブリックネットワークに接続できます。複数のパブリックネットワークインタフェースカードを持つホストを設定することによって、複数のカードをアクティブにし、それぞれを相互のフェイルオーバーバックアップとすることができます。アダプタの 1 つに障害が発生すると、Sun Cluster の Solaris IPMP ソフトウェアが呼び出され、障害のあるインタフェースが同じグループの別のアダプタにフェイルオーバーされます。

索引