



# Sun Cluster 3.0 の概念

---

Sun Microsystems, Inc.  
901 San Antonio Road  
Palo Alto, CA 94303-4900  
U.S.A. 650-960-1300

Part Number 806-6719  
2000 年 12 月, Revision A

Copyright Copyright 2000 Sun Microsystems, Inc. 901 San Antonio Road, Palo Alto, California 94303-4900 U.S.A. All rights reserved.

本製品およびそれに関連する文書は著作権法により保護されており、その使用、複製、頒布および逆コンパイルを制限するライセンスのもとにおいて頒布されます。サン・マイクロシステムズ株式会社による事前の許可なく、本製品および関連する文書のいかなる部分も、いかなる方法によっても複製することが禁じられます。

本製品の一部は、カリフォルニア大学からライセンスされている Berkeley BSD システムに基づいていることがあります。UNIX は、X/Open Company, Ltd. が独占的にライセンスしている米国ならびに他の国における登録商標です。フォント技術を含む第三者のソフトウェアは、著作権により保護されており、提供者からライセンスを受けているものです。Netscape Communicator™ は、次の著作権で保護されています。(c) Copyright 1995 Netscape Communications Corporation. All rights reserved.

RESTRICTED RIGHTS: Use, duplication, or disclosure by the U.S. Government is subject to restrictions of FAR 52.227-14(g)(2)(6/87) and FAR 52.227-19(6/87), or DFAR 252.227-7015(b)(6/95) and DFAR 227.7202-3(a).

本製品に含まれる HG 明朝 L と HG ゴシック B は、株式会社リコーがリコービマジクス株式会社からライセンス供与されたタイプフェイスマスクをもとに作成されたものです。平成明朝体 W3 は、株式会社リコーが財団法人 日本規格協会 文字フォント開発・普及センターからライセンス供与されたタイプフェイスマスクをもとに作成されたものです。また、HG 明朝 L と HG ゴシック B の補助漢字部分は、平成明朝体 W3 の補助漢字を使用しています。なお、フォントとして無断複製することは禁止されています。

Sun、Sun Microsystems、AnswerBook2、docs.sun.com、Sun Management Center、Solstice DiskSuite、Sun StorEdge は、米国およびその他の国における米国 Sun Microsystems, Inc. (以下、米国 Sun Microsystems 社とします) の商標もしくは登録商標です。

サン・のロゴマークおよび Solaris は、米国 Sun Microsystems 社の登録商標です。

すべての SPARC 商標は、米国 SPARC International, Inc. のライセンスを受けて使用している同社の米国およびその他の国における商標または登録商標です。SPARC 商標が付いた製品は、米国 Sun Microsystems 社が開発したアーキテクチャに基づくものです。

OPENLOOK、OpenBoot、JLE は、サン・マイクロシステムズ株式会社の登録商標です。

Wnn は、京都大学、株式会社アステック、オムロン株式会社で共同開発されたソフトウェアです。

Wnn6 は、オムロン株式会社で開発されたソフトウェアです。(Copyright OMRON Co., Ltd. 1999 All Rights Reserved.)

ATOK は、株式会社ジャストシステムの登録商標です。

ATOK8 は株式会社ジャストシステムの著作物であり、ATOK8 にかかる著作権その他の権利は、すべて株式会社ジャストシステムに帰属します。

ATOK Server/ATOK12 は、株式会社ジャストシステムの著作物であり、「ATOK Server/ATOK12」にかかる著作権その他の権利は、株式会社ジャストシステムおよび各権利者に帰属します。

Unicode は、Unicode, Inc. の商標です。

本書で参照されている製品やサービスに関しては、該当する会社または組織に直接お問い合わせください。

OPEN LOOK および Sun Graphical User Interface は、米国 Sun Microsystems 社が自社のユーザおよびライセンス実施権者向けに開発しました。米国 Sun Microsystems 社は、コンピュータ産業用のビジュアルまたはグラフィカル・ユーザインタフェースの概念の研究開発における米国 Xerox 社の先駆者としての成果を認めるものです。米国 Sun Microsystems 社は米国 Xerox 社から Xerox Graphical User Interface の非独占的ライセンスを取得しており、このライセンスは米国 Sun Microsystems 社のライセンス実施権者にも適用されます。

DtComboBox ウィジェットと DtSpinBox ウィジェットのプログラムおよびドキュメントは、Interleaf, Inc. から提供されたものです。(Copyright (c) 1993 Interleaf, Inc.)

Federal Acquisitions: Commercial Software—Government Users Subject to Standard License Terms and Conditions.

本書は、「現状のまま」をベースとして提供され、商品性、特定目的への適合性または第三者の権利の非侵害の黙示の保証を含みそれに限定されない、明示的であるか黙示的であるかを問わない、なんらの保証も行われないものとします。

本製品が、外国為替および外国貿易管理法 (外為法) に定められる戦略物資等 (貨物または役務) に該当する場合、本製品を輸出または日本国外へ持ち出す際には、サン・マイクロシステムズ株式会社の事前の書面による承諾を得ることのほか、外為法および関連法規に基づく輸出手続き、また場合によっては、米国商務省または米国所轄官庁の許可を得ることが必要です。

原典: Sun Cluster 3.0 Concepts

Part No: 806-1424-10

Revision A

© 2000 by Sun Microsystems, Inc.



# 目次

---

- はじめに 7
- 1. 基本知識と概要 13
  - Sun Cluster の基本知識 13
    - Sun Cluster の高可用性 14
    - Sun Cluster のフェイルオーバーとスケーラビリティ 15
  - Sun Cluster の概要 16
    - ハードウェア保守担当者 16
    - システム管理者 17
    - アプリケーションプログラマ 19
  - Sun Cluster の作業 20
- 2. 重要な概念 – ハードウェアコンポーネント 23
  - Sun Cluster ハードウェアコンポーネント 23
    - クラスタノード 24
    - 多重ホストディスク 27
    - ローカルディスク 28
    - リムーバブルメディア 29
    - クラスタインターコネクト 29
    - パブリックネットワークインタフェース 30
    - クライアントシステム 30

	管理コンソール	31
	コンソールアクセスデバイス	31
	Sun Cluster トポロジ	32
	クラスタペアトポロジ	32
	ペア +M トポロジ	33
	N+1 (星型) トポロジ	34
3.	重要な概念 – 管理とアプリケーション開発	37
	クラスタ管理とアプリケーション開発	38
	管理インタフェース	39
	クラスタ内の時間	40
	高可用性フレームワーク	41
	広域デバイス	44
	ディスクデバイスグループ	45
	広域名前空間	47
	クラスタファイルシステム	49
	定足数と定足数デバイス	52
	ボリューム管理ソフトウェア	57
	データサービス	58
	新しいデータサービスの開発	65
	リソースとリソースタイプ	67
	パブリックネットワーク管理 (PNM) とネットワークアダプタフェイ ルオーバー (NAFO)	69
4.	頻繁に寄せられる質問 (FAQ)	73
	高可用性に関する FAQ	73
	ファイルシステムに関する FAQ	74
	ボリューム管理に関する FAQ	75
	データサービスに関する FAQ	75
	パブリックネットワークに関する FAQ	76

クラスタメンバーに関する FAQ	77
クラスタ記憶装置に関する FAQ	78
クラスタインターコネクトに関する FAQ	78
クライアントシステムに関する FAQ	79
管理コンソールに関する FAQ	79
端末集配信装置とシステムサービスプロセッサに関する FAQ	80
用語集	83



## はじめに

---

このマニュアルでは、Sun Cluster ソフトウェアの概念と参照情報について説明します。

このマニュアルは、Sun のソフトウェアとハードウェアについて幅広い知識を持っている上級システム管理者を対象としています。販売活動のガイドとしては使用しないでください。このマニュアルを読む前に、システムの必要条件を確認し、適切な装置とソフトウェアを購入しておく必要があります。

このマニュアルで説明されている作業手順を行うには、Solaris™ オペレーティング環境に関する知識と、Sun Cluster と共に使用するボリューム管理ソフトウェアに関する専門知識が必要になります。

---

## 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%
<b>AaBbCc123</b>	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% <b>su</b>  password:
<i>AaBbCc123</i>	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 <i>filename</i> 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。
\	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% <b>grep</b> `^#define \ XV_VERSION_STRING`

ただし AnswerBook2 では、ユーザーが入力する文字と画面上のコンピュータ出力は区別して表示されません。

コード例は次のように表示されます。

■ C シェルプロンプト

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

■ Bourne シェルおよび Korn シェルのプロンプト

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

■ スーパーユーザーのプロンプト



```
system# 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。

---

## 関連マニュアル

説明内容	タイトル	Part No.
インストール	『 <i>Sun Cluster 3.0</i> ソフトウェアのインストール』	806-6725
ハードウェア	『 <i>Sun Cluster 3.0 Hardware Guide</i> 』	806-1420
データサービス	『 <i>Sun Cluster 3.0</i> データサービスのインストールと構成』	806-6729
API 開発	『 <i>Sun Cluster 3.0</i> データサービス開発ガイド』	806-6723
管理	『 <i>Sun Cluster 3.0</i> のシステム管理』	806-6731
エラーメッセージと問題解決	『 <i>Sun Cluster 3.0 Error Messages Manual</i> 』	806-1426
最新情報	『 <i>Sun Cluster 3.0</i> ご使用にあたって』	806-6735

---

## Sun のマニュアルの注文方法

専門書を扱うインターネットの書店 Fatbrain.com から、米国 Sun Microsystems™, Inc. (以降、Sun™ とします) のマニュアルをご注文いただけます。

マニュアルのリストと注文方法については、<http://www1.fatbrain.com/documentation/sun> の Sun Documentation Center をご覧ください。

---

## Sun のオンラインマニュアル

<http://docs.sun.com> では、Sun が提供しているオンラインマニュアルを参照することができます。マニュアルのタイトルや特定の主題などをキーワードとして、検索を行うこともできます。

---

## 問い合わせについて

Sun Cluster のインストールまたは使用で問題が発生した場合は、ご購入先に連絡し、次の情報をお伝えください。

- 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
- ご使用の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バージョン番号 (例: Solaris 8)
- Sun Cluster のバージョン番号 (例: Sun Cluster 3.0)

システムの各ノードに関する情報を収集するには、次のコマンドを使用してください。

---

コマンド	機能
<code>prtconf -v</code>	システムメモリーのサイズと周辺デバイス情報を表示する
<code>psrinfo -v</code>	プロセッサの情報を表示する

---

---

コマンド	機能
<code>showrev --p</code>	インストールされているパッチを報告する
<code>prtdiag -v</code>	システム診断情報を表示する
<code>scinstall -pv</code>	Sun Cluster 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

---

`/var/adm/messages` ファイルの内容もご購入先にお知らせください。



## 基本知識と概要

---

このマニュアルでは、Sun Cluster のマニュアルの読者に必要な概念について説明します。次の読者を対象としています。

- クラスタハードウェアを設置して保守を行う担当者
- Sun Cluster ソフトウェアをインストール、構成、管理するシステム管理者
- 現在 Sun Cluster 製品に含まれていないアプリケーション用のデータサービスを開発するアプリケーション開発者

このマニュアルは、Sun Cluster の他のマニュアルと合わせて、Sun Cluster の全体を説明するものです。

この章では、次のことを説明します。

- Sun Cluster の基本知識と概要
- Sun Cluster の各ユーザーごとの役割と参照する情報
- Sun Cluster で作業するにあたって理解する必要がある重要な概念
- 重要な概念に関連する手順と情報を記載した Sun Cluster のマニュアル
- クラスタに関連する作業と、これらの作業手順が記載されたマニュアル

---

## Sun Cluster の基本知識

Sun Cluster は、Solaris™ オペレーティング環境をクラスタオペレーティングシステムに拡張するものです。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、1つのクライアントとして扱います。

各クラスタノードは、それ自身のプロセスを実行するスタンドアロンサーバーです。これらのプロセスは相互にやりとりして、ユーザーに提供するアプリケーション、システムリソース、データを(ネットワーククライアントに対して)1つのシステムとして見えるように形成します。

クラスタは、従来の単一サーバーシステムと比較した場合、いくつかの利点があります。これらの利点には、可用性の高いスケーラブルなアプリケーションのサポート、モジュールの成長に対応できる容量、従来のハードウェアフォルトトレラントシステムよりも低価格の製品といったものがあります。

次に、Sun Cluster の導入目的を示します。

- ソフトウェアまたはハードウェアの障害が原因のシステム停止時間を減らす、または完全になくす
- 単一サーバーシステムを停止させるような障害が発生しても、エンドユーザーへのデータとアプリケーションの可用性を保証する
- クラスタにノードを追加し、追加したプロセッサに応じたサービスを提供できるようにすることで、アプリケーションのスループットを向上させる
- クラスタ全体を停止しなくても保守を実行できるようにすることで、システムの可用性を強化する

## Sun Cluster の高可用性

Sun Cluster は、高可用性 (HA) システムとして設計されています。つまり、データとアプリケーションに対してほぼ連続的なアクセスを可能にするシステムです。

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。また、通常はソフトウェアの障害を考慮していません。

Sun Cluster は、ハードウェアとソフトウェアの組み合わせによって高可用性を実現しています。冗長なクラスタインターコネクト、記憶装置、パブリックネットワークは、単一の障害に対する防護策となります。クラスタソフトウェアは、メンバーノードの状態を常に監視し、障害が発生したノードがクラスタに属さないようにしてデータの破壊を防止します。また、クラスタは、アプリケーションとそれに依存

するシステムリソースを監視し、障害が発生した場合にアプリケーションの処理を継続するか再開します。

高可用性については、73ページの「高可用性に関する FAQ」を参照してください。

## Sun Cluster のフェイルオーバーとスケーラビリティ

Sun Cluster を使用すると、フェイルオーバーまたはスケーラブルのどちらかをベースにしてアプリケーションを実装できます。フェイルオーバーおよびスケーラブルの各アプリケーションは、同じクラスタで同時に実行することもできます。通常、フェイルオーバーアプリケーションは可用性 (冗長性) が高く、スケーラブルアプリケーションは可用性が高いとともに、パフォーマンスも向上します。

### フェイルオーバー

フェイルオーバーとは、クラスタが、障害の発生した主ノードから指定した二次ノードにアプリケーションを自動的に再配置するプロセスのことです。フェイルオーバーによって、Sun Cluster は高い可用性を実現します。

フェイルオーバーが発生すると、クライアントでは、サービスが短時間中断して、フェイルオーバーの終了後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、アプリケーションとデータの提供元である物理サーバーを認識していません。

### スケーラビリティ

フェイルオーバーは冗長性に関係していますが、スケーラビリティは負荷に関係なく一定した応答時間とスループットを提供するものです。スケーラブルアプリケーションは、1つのクラスタにある複数のノードに作用し、アプリケーションを同時に実行するため、パフォーマンスは向上します。スケーラブルな構成では、クラスタ内の各ノードは、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。

フェイルオーバーとスケーラブルサービスの詳細については、58ページの「データサービス」を参照してください。

---

## Sun Cluster の概要

この節では、Sun Cluster のユーザーを 3 種類に分け、各ユーザーに関連する概念とマニュアルについて説明します。各ユーザーは次のとおりです。

- ハードウェア保守担当者
- システム管理者
- アプリケーションプログラマ。Sun Cluster は、一連の可用性の高いデータサービスを提供します。これらのサービスは、クラスタで実行される可用性の高いデータサービスとして構成された Oracle、Apache Web Server、DNS などのアプリケーションです。Sun Cluster API を使用すると、可用性の高いデータサービスに他のアプリケーションを組み込むことができます。アプリケーションプログラマは、API を使用するシェルスクリプトや C プログラムを作成できます。

### ハードウェア保守担当者

ハードウェア保守担当者にとって、Sun Cluster は、サーバー、ネットワーク、および記憶装置を含む市販のハードウェアの集合に見えます。これらのコンポーネントは、すべてのコンポーネントにバックアップがあり、単一の障害によってシステム全体が停止しないように配線されています。

### 重要な概念 (ハードウェア保守担当者)

ハードウェア保守担当者は、クラスタに関する次の概念を理解する必要があります。

- クラスタハードウェアの構成と配線
- 設置と保守 (追加、取り外し、交換)
  - ネットワークインタフェースコンポーネント (アダプタ、接続点、ケーブル)
  - ディスクインタフェースカード
  - ディスクアレイ
  - ディスクドライブ
  - 管理コンソールとコンソールアクセスデバイス
- 管理コンソールとコンソールアクセスデバイスの設定



## 参照箇所 (ハードウェア保守担当者)

次の項には、前述の重要な概念に関連する説明が記載されています。

- 24ページの「クラスタノード」
- 27ページの「多重ホストディスク」
- 28ページの「ローカルディスク」
- 29ページの「クラスタインターコネクト」
- 30ページの「パブリックネットワークインタフェース」
- 30ページの「クライアントシステム」
- 31ページの「管理コンソール」
- 31ページの「コンソールアクセスデバイス」
- 32ページの「クラスタペアトポロジ」
- 33ページの「ペア +M トポロジ」
- 34ページの「N+1 (星型) トポロジ」

## Sun Cluster の関連マニュアル (ハードウェア保守担当者)

次の Sun Cluster のマニュアルには、ハードウェア保守の概念に関連する手順と情報が記載されています。

- 『*Sun Cluster 3.0 Hardware Guide*』

## システム管理者

システム管理者にとって、Sun Cluster は、ケーブルによって接続された、記憶装置を共有するサーバー (ノード) の集合に見えます。システム管理者は、次のソフトウェアを扱います。

- クラスタノード間のコネクティビティを監視するための、Solaris ソフトウェアに統合された専用のクラスタソフトウェア
- クラスタノードで実行されるユーザーアプリケーションプログラムの状態を監視するための専用のソフトウェア
- ディスクを設定して管理するためのボリューム管理ソフトウェア
- 直接ディスクに接続されていないものも含め、すべてのノードが、すべての記憶装置にアクセスできるようにするための専用のクラスタソフトウェア

- ファイルがすべてのノードに対してローカルに接続されているように表示するための専用のソフトウェア

## 重要な概念 (システム管理者)

システム管理者は、次の概念とプロセスについて理解する必要があります。

- ハードウェアとソフトウェアの間の対話
- クラスタをインストールして構成する方法の一般的な流れ
  - Solaris オペレーティング環境のインストール
  - Sun Cluster のインストールと構成
  - ボリューム管理ソフトウェアのインストールと構成
  - クラスタを動作可能状態にするためのアプリケーションソフトウェアのインストールと構成
  - Sun Cluster データサービスソフトウェアのインストールと構成
- クラスタハードウェアとソフトウェアのコンポーネントを追加、削除、交換、およびサービス提供するためのクラスタ管理手順
- パフォーマンスを向上させるための構成の変更方法

## 参照箇所 (システム管理者)

次の項には、前述の主な概念に関連する説明が記載されています。

- 39ページの「管理インタフェース」
- 41ページの「高可用性フレームワーク」
- 44ページの「広域デバイス」
- 45ページの「ディスクデバイスグループ」
- 47ページの「広域名前空間」
- 49ページの「クラスタファイルシステム」
- 52ページの「定足数と定足数デバイス」
- 57ページの「ボリューム管理ソフトウェア」
- 58ページの「データサービス」
- 67ページの「リソースとリソースタイプ」

- 69ページの「パブリックネットワーク管理 (PNM) とネットワークアダプタ フェイルオーバー (NAFO)」
- 第4章

## Sun Cluster の関連マニュアル (システム管理者)

次の Sun Cluster のマニュアルには、システム管理の概念に関連する手順と情報が記載されています。

- 『Sun Cluster 3.0 ソフトウェアのインストール』
- 『Sun Cluster 3.0 のシステム管理』
- 『Sun Cluster 3.0 Error Messages Manual』

## アプリケーションプログラマ

Sun Cluster は、Oracle、NFS、DNS、iPlanet Web Server、Apache Web Server、Netscape Directory Server などのアプリケーションに対して、可用性の高いデータサービスを提供します。あるサイトで、別のアプリケーションをクラスタ上で実行する必要が生じた場合は、Sun Cluster API (アプリケーションプログラミングインタフェース) と DSDL API (データサービス開発ライブラリ API) を使用して、そのアプリケーションをクラスタ上で可用性の高いデータサービスとして実行するために必要なデータサービスソフトウェアを開発できます。

## 重要な概念 (アプリケーションプログラマ)

アプリケーションプログラマは、次の点について理解する必要があります。

- 各アプリケーションの特性。アプリケーションを可用性の高いデータサービスまたはスケーラブルなデータサービスとして実行できるかどうかを判断する必要があります。
- Sun Cluster API、DSDL API、汎用データサービス。プログラマは、各自のアプリケーションをクラスタ環境に合わせて構成するプログラムまたはスクリプトを記述するために、どのツールが最も適しているかを判断する必要があります。

## 参照箇所 (アプリケーションプログラマ)

次の項には、前述の主な概念に関連する説明が記載されています。

- 58ページの「データサービス」
- 67ページの「リソースとリソースタイプ」
- 第4章

## Sun Cluster の関連マニュアル (アプリケーションプログラマ)

次の Sun Cluster のマニュアルには、アプリケーションプログラミングの概念に関連する手順と情報が記載されています。

- 『Sun Cluster 3.0 データサービス開発ガイド』
- 『Sun Cluster 3.0 データサービスのインストールと構成』

---

## Sun Cluster の作業

すべての概念は作業に対応し、すべての作業にいくつかの概念的な予備知識が必要です。次の表に、作業と作業手順を記載しているマニュアルを示します。このマニュアルの概念に関する章では、各概念がこれらの作業とどのように対応するかを説明します。

表 1-1 作業マップ: ユーザーの作業と参照するマニュアル

実行する作業	使用するマニュアル
クラスタハードウェアの設置	『Sun Cluster 3.0 Hardware Guide』
クラスタへの Solaris ソフトウェアのインストール	『Sun Cluster 3.0 ソフトウェアのインストール』
Sun <sup>TM</sup> Management Center ソフトウェアのインストール	『Sun Cluster 3.0 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と構成	『Sun Cluster 3.0 ソフトウェアのインストール』
ボリューム管理ソフトウェアのインストールと構成	『Sun Cluster 3.0 ソフトウェアのインストール』 各ボリューム管理ソフトウェアのマニュアル
Sun Cluster データサービスのインストールと構成	『Sun Cluster 3.0 データサービスのインストールと構成』

表 1-1 作業マップ: ユーザーの作業と参照するマニュアル 続く

実行する作業	使用するマニュアル
クラスタハードウェアの保守	『Sun Cluster 3.0 Hardware Guide』
Sun Cluster ソフトウェアの管理	『Sun Cluster 3.0 のシステム管理』
ボリューム管理ソフトウェアの管理	『Sun Cluster 3.0 のシステム管理』 各ボリューム管理ソフトウェアのマニュアル
アプリケーションソフトウェアの管理	各アプリケーションのマニュアル
問題の識別と対処方法	『Sun Cluster 3.0 Error Messages Manual』
新しいデータサービスの作成	『Sun Cluster 3.0 データサービス開発ガイド』



## 重要な概念 – ハードウェアコンポーネント

---

この章では、Sun Cluster 構成のハードウェアコンポーネントに関連する重要な概念について説明します。

---

### Sun Cluster ハードウェアコンポーネント

ここで示す情報は、主にハードウェアサービスプロバイダを対象としています。これらの概念は、サービスプロバイダが、クラスタハードウェアの設置、構成、またはサービスを提供する前に、ハードウェアコンポーネント間の関係を理解するのに役立ちます。またこれらの情報は、クラスタシステムの管理者にとっても、クラスタソフトウェアをインストール、構成、管理するための予備知識として役立ちます。

クラスタは、次のようなハードウェアコンポーネントで構成されます。

- ローカルディスクを備えたクラスタノード (非共有)
- 多重ホスト記憶装置 (ノード間で共有されるディスク)
- リムーバブルメディア (テープ、CD-ROM)
- クラスタインターコネクタ
- 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
- 管理コンソール
- コンソールアクセスデバイス

Sun Cluster を使用すると、32ページの「Sun Cluster トポロジ」で説明するように、これらのコンポーネントを各種の構成に組み合わせることができます。

次の図は、クラスタの構成例を示しています。

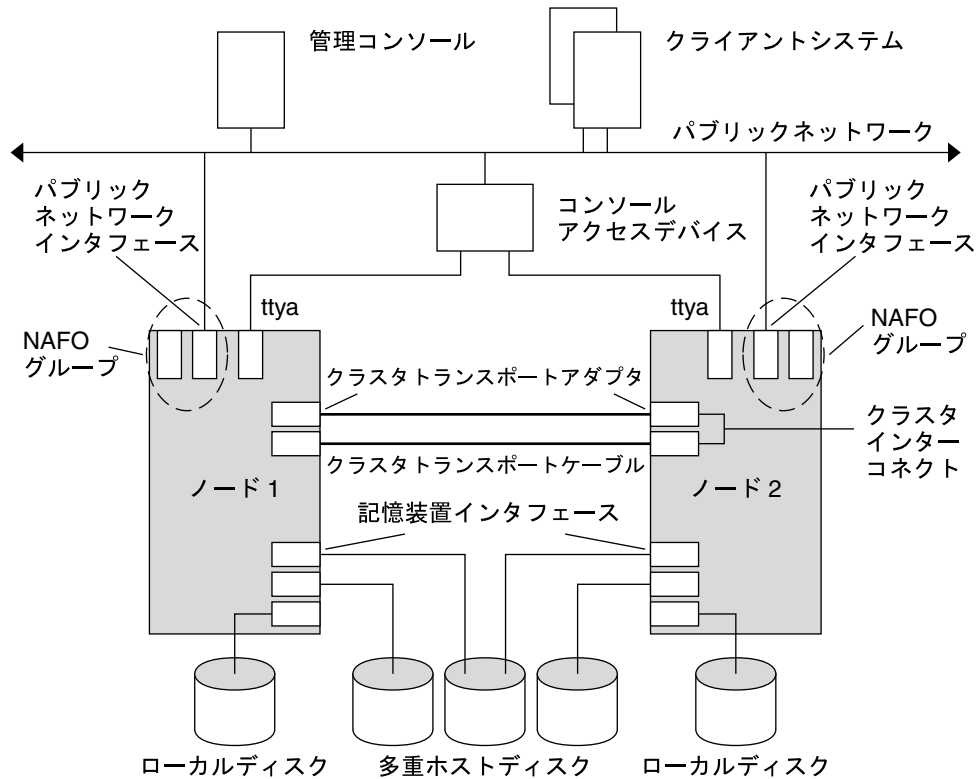


図 2-1 2 ノードクラスタ構成の例

## クラスタノード

クラスタノードとは、Solaris オペレーティング環境と Sun Cluster ソフトウェアの両方を実行するマシンのことで、クラスタの現在のメンバー (クラスタメンバー) または潜在的なメンバーのどちらかです。Sun Cluster ソフトウェアを使用すると、1つのクラスタに2～8台のノードを設定できます。サポートされるノード構成については、32ページの「Sun Cluster トポロジ」を参照してください。

一般的にクラスタノードは、1つまたは複数の多重ホストディスクに接続されます。スケーラブルサービスを1つ構成することで、ノードが多重ホストディスクに直接接続されていなくてもサービスを提供することができます。多重ホストディス



クに接続されていないノードは、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。

パラレルデータベース構成では、複数のノードがすべてのディスクへの同時アクセスを共有します。パラレルデータベース構成については、27ページの「多重ホストディスク」と第3章を参照してください。

クラスタ内のノードはすべて、共通の名前(クラスタ名)によってグループ化されます。この名前は、クラスタのアクセスと管理に使用されます。

パブリックネットワークアダプタは、ノードとパブリックネットワークを接続して、クラスタへのクライアントアクセスを可能にします。

クラスタメンバーは、プライベートネットワークと呼ばれる1つまたは複数の物理的に独立したネットワークを介して、クラスタ内の他のノードとやりとりします。クラスタ内のプライベートネットワークの集合は、クラスタインターコネクトと呼ばれます。

クラスタ内のすべてのノードが、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、他のクラスタノードで実行されているリソースだけでなく、ローカルに実行されているリソースも認識します。

クラスタメンバーとリソース(アプリケーション、ディスク記憶装置など)は、フェイルオーバーやスケラブル機能を実現できる形式で構成します。

同じクラスタ内の各ノードの処理、メモリー、および入出力機能が同等のもので、パフォーマンスを著しく低下させることなく処理を継続できることを確認してください。フェイルオーバーの可能性があるため、すべてのノードに、バックアップまたは二次ノードとしてすべてのノードの作業負荷を引き受けるのに十分な予備容量があることを確認してください。

各ノードは、独自のルート(/)ファイルシステムを起動します。

## クラスタメンバー用のソフトウェアコンポーネント

クラスタメンバーとして機能するには、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 オペレーティング環境
- Sun Cluster
- ボリューム管理ソフトウェア (Solstice DiskSuite™ または VERITAS Volume Manager)

■ データサービスアプリケーション

例外として、複数のディスクの冗長配列 (RAID) を使用する Oracle Parallel Sever(OPS) 構成があります。この構成には、Oracle データを管理するために Solstice DiskSuite や VERITAS Volume Manager などのボリューム管理ソフトウェアは必要ありません。

Solaris オペレーティング環境、Sun Cluster、およびボリューム管理ソフトウェアをインストールする方法については、『Sun Cluster 3.0 ソフトウェアのインストール』を参照してください。データサービスをインストールして構成する方法については、『Sun Cluster 3.0 データサービスのインストールと構成』を参照してください。

前述のソフトウェアコンポーネントの概念については、第 3 章を参照してください。

次の図は、Sun Cluster ソフトウェア環境を作成するために共に動作するソフトウェアコンポーネントの概要を示しています。

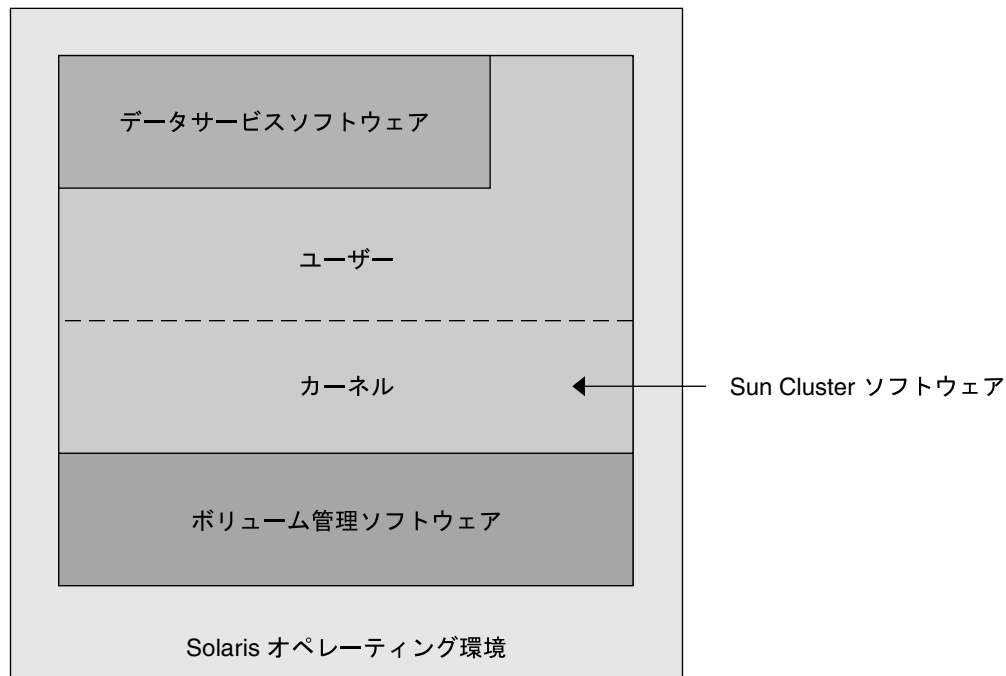


図 2-2 Sun Cluster ソフトウェアコンポーネントの関係の概要

クラスタメンバーについては、第 4 章を参照してください。

## 多重ホストディスク

Sun Cluster には、一度に複数のノードに接続可能なディスクである、多重ホストディスク記憶装置が必要です。Sun Cluster 環境では、多重ホスト記憶装置によってディスクデバイスの可用性を強化できます。多重ホスト記憶装置に常駐するディスクデバイスは、単一ノードの障害に耐えられます。

多重ホストディスクはアプリケーションデータを保存するだけでなく、データサービスバイナリと構成ファイルも保存できます。

多重ホストディスクは、ディスクのマスターとなる主ノードによって広域的にアクセスされるか、ローカルパスによって同時に直接アクセスされます。現在、直接同時アクセスを使用するアプリケーションは OPS だけです。

多重ホストディスクはノード障害を防止します。クライアント要求があるノードを介してデータにアクセスしていて失敗した場合、これらの要求は、同じディスクへの直接接続を持つ別のノードを使用するようにスイッチオーバーされます。

ボリューム管理ソフトウェアは、多重ホストディスクのデータ冗長性に対して、ミラー化された構成または RAID-5 構成を提供します。現在、Sun Cluster は、Solstice DiskSuite と VERITAS Volume Manager をボリューム管理ソフトウェアとして、また RDAC RAID-5 ハードウェアコントローラを Sun StorEdge™ A3x00 記憶ユニット用にサポートしています。

多重ホストディスクをミラー化およびストライプ化したディスクと組み合わせると、ノードの障害および個々のディスクの障害の両方に対する防御策となります。

多重ホスト記憶装置については、第 4 章を参照してください。

## 多重イニシエータ SCSI

この項は、多重ホストディスクに使用されるファイバチャネル記憶装置ではなく、SCSI 記憶装置にのみ適用されます。

スタンドアロンサーバーでは、サーバーノードが、このサーバーを特定の SCSI バスに接続する SCSI ホストアダプタ回路によって、SCSI バスのアクティビティを制御します。この SCSI ホストアダプタ回路は、SCSI イニシエータと呼ばれます。この回路は、この SCSI バスに対するすべてのバスアクティビティを開始します。Sun システムの SCSI ホストアダプタのデフォルト SCSI アドレスは 7 です。

クラスタ構成は、複数のサーバーノード間で記憶装置を共有します。クラスタ記憶装置が SCSI デバイスまたは Differential SCSI デバイスからなる場合、この構成は多

重イニシエータ SCSI と呼ばれます。この用語が示すように、複数の SCSI イニシエータが SCSI バスに存在します。

SCSI 仕様では、SCSI バスの各デバイスに一意の SCSI アドレスが必要です。(ホストアダプタも、SCSI バス上のデバイスの 1 つです。) 多重イニシエータ環境でデフォルトのハードウェア構成を行うと、すべての SCSI ホストアダプタがデフォルトで 7 に設定されるため衝突が生じます。

この衝突を解決するには、各 SCSI バスで、SCSI アドレスが 7 の SCSI ホストアダプタを 1 つ残し、他のホストアダプタには、未使用の SCSI アドレスを設定します。これらの未使用の SCSI アドレスには、現在未使用のアドレスと最終的に未使用となるアドレスの両方を含めるべきです。将来未使用となるアドレスの例としては、新しいドライブを空のドライブスロットに設置することによる記憶装置の追加があります。ほとんどの構成では、二次ホストアダプタに使用できる SCSI アドレスは 6 です。

これらのホストアダプタに指定された SCSI アドレスは、`scsi-initiator-id` Open Boot PROM (OBP) プロパティを設定することにより変更できます。このプロパティは 1 つのノードに対して、またはホストアダプタごとに広域的に設定できます。一意の `scsi-initiator-id` を各 SCSI ホストアダプタに設定するための手順は、『*Sun Cluster 3.0 Hardware Guide*』の各ディスク格納装置に関する章に記載されています。

## ローカルディスク

ローカルディスクとは、単一ノードにのみ接続されたディスクをいいます。したがって、これらはノードの障害に対して保護されていません(可用性が高くありません)。ただし、ローカルディスクを含むすべてのディスクが広域的名前空間に含まれ、広域デバイスとして構成されています。したがって、ディスク自体をすべてのクラスタノードから参照できます。これらのディスク上のファイルシステムは、広域マウントポイントに入れることによって、他のノードで使用できるようにすることができます。これらの広域ファイルシステムのいずれかがマウントされているノードに障害が生じると、すべてのノードがそのファイルシステムにアクセスできなくなります。ボリューム管理ソフトウェアを使用すると、これらのディスクをミラー化することができ、ディスク障害によってこれらのファイルシステムがアクセス不能になるのを防止できますが、ボリューム管理ソフトウェアはノード障害を防御できません。

## リムーバブルメディア

クラスタでは、テープドライブや CD-ROM ドライブなどのリムーバブルメディアがサポートされています。通常、これらのデバイスは、非クラスタ化環境と同じ方法で設置して構成することで使用できます。これらのデバイスは、Sun Cluster で広域デバイスとして構成されるため、クラスタ内の任意のノードから各デバイスにアクセスできます。リムーバブルメディアの設置方法と構成については、『Sun Cluster 3.0 Hardware Guide』を参照してください。

## クラスタインターコネクト

クラスタインターコネクトとは、クラスタプライベート通信とデータサービス通信をクラスタノード間で転送するために使用される、デバイスの物理構成のことです。インターコネクトは、クラスタプライベート通信で拡張使用されるため、パフォーマンスが制限される可能性があります。

クラスタノードだけがプライベートインターコネクトに接続できます。Sun Cluster セキュリティモデルは、クラスタノードだけがプライベートインターコネクトに物理的にアクセスできるものと想定しています。

少なくとも 2 つの冗長なプライベートネットワーク、またはパスを使用して、すべてのノードをクラスタインターコネクトによって接続し、単一地点による障害を回避する必要があります。任意の 2 つのノード間で、複数のプライベートネットワーク (2 ~ 6) を設定できます。クラスタインターコネクトは、アダプタ、接続点、およびケーブルの 3 つのハードウェアコンポーネントからなります。各プライベートネットワークは、他のプライベートネットワークと共通のハードウェアコンポーネントを共有しないように構成されています。

次に、これらの各ハードウェアコンポーネントを説明します。

- アダプタ – 各クラスタノードに常駐する物理ネットワークカード。この名前は、たとえば、Quad FastEthernet を示す qfe など、製品の名前からとられます。一部のアダプタには物理ネットワーク接続が 1 つしかありませんが、qfe カードのように複数の物理接続を持つものもあります。ネットワークインタフェースと記憶装置インタフェースの両方を持つものもあります。

複数のインタフェースを持つネットワークカードは、カード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。可用性を最適にするには、2 つのノード間の唯一のパスが、単一のネットワークカードに依存しないように、クラスタを設定してください。

- 接続点 - クラスタノードの外部に常駐するスイッチ。これらは、パススルーおよび切り換え機能を実行して、3つ以上のノードに同時に接続できるようにします。2ノードクラスタでは、各ノードの冗長アダプタに接続された冗長物理ケーブルによって、ノードを相互に直接接続できるため、接続点は必要ありません。3ノード以上の構成では、通常は接続点が必要です。
- 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接続点の間をつなぐ物理接続。

クラスタインターコネクトについては、第4章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。ノードは、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構成して、1つのカードがアクティブで、もう1つのカードがバックアップとして動作するように設定できます。Sun Cluster ソフトウェアのサブシステムであるパブリックネットワーク管理 (PNM) は、アクティブインタフェースを監視します。アクティブアダプタに障害が生じると、ネットワークアダプタフェイルオーバー (NAFO) ソフトウェアが呼び出されて、バックアップアダプタの1つにインタフェースの処理を継続します。

パブリックネットワークインタフェースのクラスタ化に関連する特殊なハードウェアの考慮事項はありません。

パブリックネットワークについては、第4章を参照してください。

## クライアントシステム

クライアントシステムには、パブリックネットワークによってクラスタにアクセスするワークステーションや他のサーバーが含まれます。クライアント側プログラムは、クライアントで実行されるサーバー側アプリケーションによって提供されるデータやサービスを使用します。

クライアントシステムの可用性は高くありません。クラスタ上のデータとアプリケーションは、高い可用性を備えています。

クライアントシステムについては、第4章を参照してください。

## 管理コンソール

管理コンソールと呼ばれる専用の SPARCstation™ システムを使用して、アクティブ クラスタを管理できます。通常、Cluster Control Panel (CCP) や、Sun Management Center 用の Sun Cluster モジュールなどの管理ツールソフトウェアを管理コンソールにインストールして実行します。CCP で `cconsole` を使用すると、一度に複数の ノードコンソールに接続できます。CCP の使用法の詳細については、『Sun Cluster 3.0 のシステム管理』を参照してください。

管理コンソールはクラスタノードではありません。管理コンソールは、パブリック ネットワークを介して、または任意でネットワークベースの端末集配信装置 (コンセントレータ) を介して、クラスタノードへの遠隔アクセス用に使用します。クラスタ が Sun™ Enterprise E10000 プラットフォームで構成されている場合は、管理コンソールからシステムサービスプロセッサ (SSP) にログインし、`netcon(1M)` コマンドを使用して接続する必要があります。

通常、ノードはモニターなしで構成します。ノードのコンソールには、端末集配信装置に接続され、さらにその端末集配信装置からノードのシリアルポートに接続された管理コンソールから `telnet` セッションによってアクセスします (Sun Enterprise E10000 サーバーの場合は、システムサービスプロセッサ (SSP) から接続します)。詳細は、31ページの「コンソールアクセスデバイス」を参照してください。

Sun Cluster には、専用の管理コンソールは必要ありませんが、専用コンソールを使用すると、次の利点が得られます。

- コンソールと管理ツールを同じマシンにグループ化することによって、クラスタの集中管理が可能になる
- ハードウェアサービスプロバイダによる問題解決が迅速に行われる

管理コンソールについては、第 4 章を参照してください。

## コンソールアクセスデバイス

すべてのクラスタノードへのコンソールアクセスが必要です。コンソールアクセスを取得するには、クラスタハードウェアとともに購入した端末集配信装置、Sun Enterprise E10000 サーバー上のシステムサービスプロセッサ (SSP)、または各ノードの `ttya` にアクセスできる他のデバイスを使用してください。

サポートされている唯一の端末集配信装置は、Sun から提供されています。サポートされている Sun の端末集配信装置の使用はオプションです。端末集配信装置を使用すると、TCP/IP ネットワークを使用して、各ノードの `ttya` にアクセスできま

す。この結果、ネットワークの任意の場所にあるリモートワークステーションから、各ノードにコンソールレベルでアクセスできます。

システムサービスプロセッサ (SSP) は、Sun Enterprise E10000 サーバーへのコンソールアクセスを提供します。SSP は、Ethernet ネットワーク上の SPARCstation であり、Sun Enterprise E10000 サーバーをサポートするように構成されています。SSP は、Sun Enterprise E10000 のサーバーの管理コンソールです。Sun Enterprise E10000 のネットワークコンソール機能を使用すると、ネットワーク上のすべてのワークステーションがホストコンソールセッションを開くことができます。

これ以外のコンソールアクセス方式には、他の端末集配信装置、別ノードおよびダンプ端末からの tip (1) シリアルポートアクセスが含まれます。Sun™ キーボードとモニター、または他のシリアルポートデバイスも使用できます。

コンソールデバイスについては、第 4 章を参照してください。

---

## Sun Cluster トポロジ

トポロジとは、クラスタノードと、クラスタで使用される記憶装置プラットフォームを接続する接続スキーマをいいます。

Sun Cluster は、次のトポロジをサポートしています。

- クラスタペア
- N+1 (星型)

次の項では、各トポロジについて説明します。

### クラスタペアトポロジ

クラスタペアトポロジとは、単一のクラスタ管理フレームワークのもとで動作する複数のノードペアをいいます。この構成では、ペアの間でのみフェイルオーバーが発生します。ただし、すべてのノードがプライベートネットワークによって接続されていて、Sun Cluster ソフトウェア制御のもとで動作します。このトポロジを使用して、1 つのペアでパラレルデータベースアプリケーションを実行し、別のペアで可用性の高いアプリケーションを実行できます。クラスタファイルシステムを使用すると、すべてのノードがアプリケーションデータを保存するディスクに直接接続されていない場合でも、複数のノードがスケラブルサービス、またはパラレルデータベースを実行する 2 ペア構成を設定できます。



次の図は、クラスタペア構成を示しています。

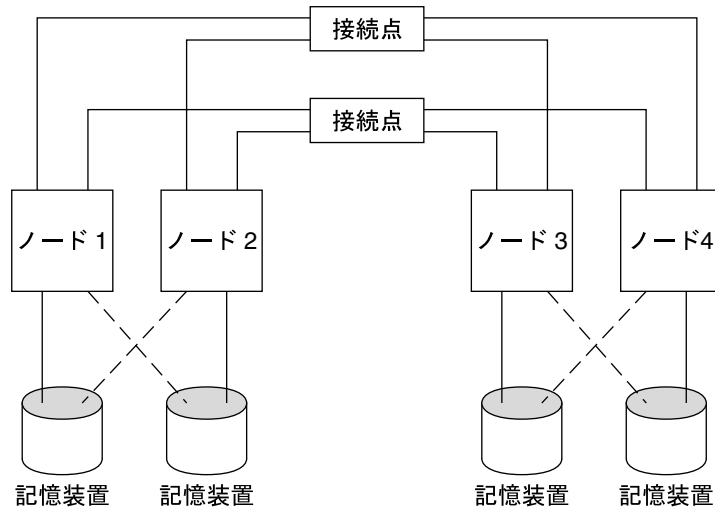


図 2-3 クラスタペアトポロジ

## ペア +M トポロジ

ペア +M トポロジには、共有記憶装置に直接接続されたノードのペアと、クラスタインターコネクトを使用して共有記憶装置にアクセスするノードの追加セットが含まれます。これらのノードはそれぞれ直接には接続されていません。この構成のノードはすべてボリューム管理ソフトウェアで構成されます。

次の図は、4つのノードのうち2つ(ノード3とノード4)がクラスタインターコネクトを使用して記憶装置にアクセスする、1つのペア +M トポロジを示しています。この構成を拡張し、共有記憶装置には直接アクセスしない追加ノードを含めることができます。

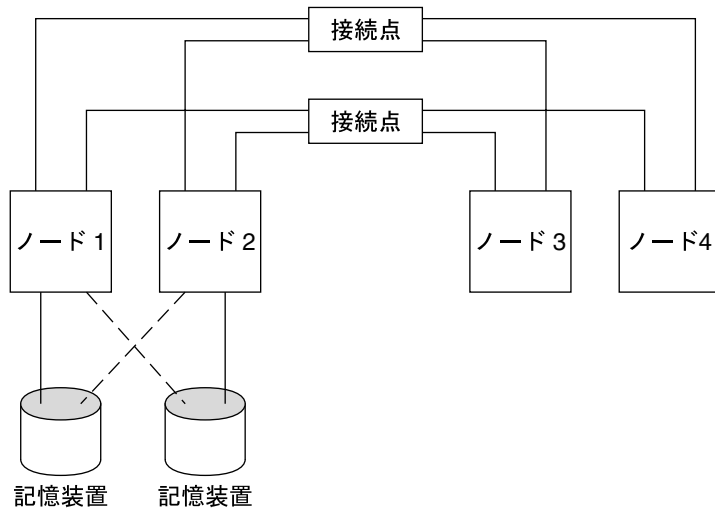


図 2-4 ペア +M トポロジ

## N+1 (星型) トポロジ

N+1 トポロジには、いくつかの主ノードと 1 つの二次ノードが含まれます。主ノードと二次ノードを同等に構成する必要はありません。主ノードは、アプリケーションサービスをアクティブに提供します。二次ノードは、主ノードに障害が生じるのを待機する間、アイドル状態である必要はありません。

二次ノードは、この構成ですべての多重ホスト記憶装置に物理的に接続されている唯一のノードです。

主ノードで障害が発生すると、Sun Cluster はそのリソースの処理を二次ノードで続行し、リソースは自動または手動で主ノードに切り換えられるまで二次ノードで機能します。

二次ノードには、主ノードの 1 つに障害が発生した場合に負荷を処理できるだけの十分な予備の CPU 容量が常に必要です。

次の図は、N+1 構成を示しています。

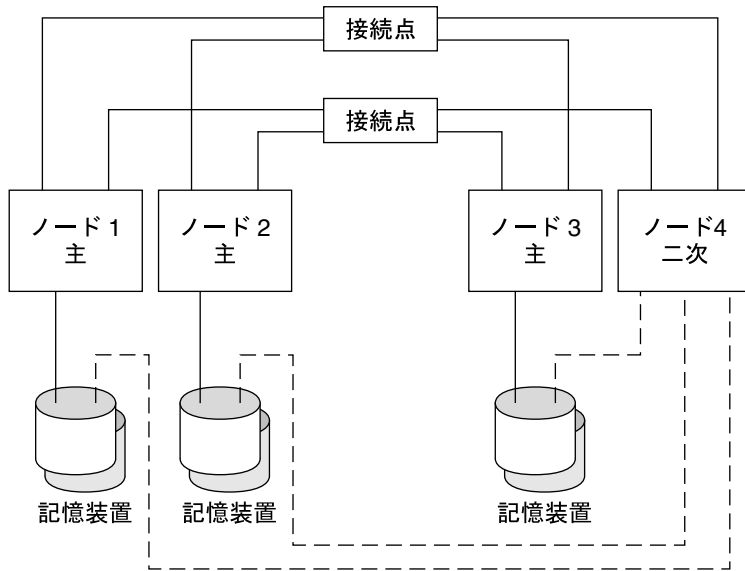


図 2-5 N+1 トポロジ



## 重要な概念 – 管理とアプリケーション開発

---

この章では、Sun Cluster 構成のソフトウェアコンポーネントに関連する重要な概念について説明します。次のトピックについて述べます。

- 39ページの「管理インタフェース」
- 40ページの「クラスタ内の時間」
- 41ページの「高可用性フレームワーク」
- 44ページの「広域デバイス」
- 45ページの「ディスクデバイスグループ」
- 47ページの「広域名前空間」
- 49ページの「クラスタファイルシステム」
- 52ページの「定足数と定足数デバイス」
- 57ページの「ボリューム管理ソフトウェア」
- 58ページの「データサービス」
- 65ページの「新しいデータサービスの開発」
- 67ページの「リソースとリソースタイプ」
- 69ページの「パブリックネットワーク管理 (PNM) とネットワークアダプタフェイルオーバー (NAFO)」

---

## クラスタ管理とアプリケーション開発

この情報は、主に、Sun Cluster API および SDK を使用するシステム管理者とアプリケーション開発者を対象としています。クラスタシステムの管理者にとっては、この情報は、クラスタソフトウェアのインストール、構成、管理についての予備知識となります。アプリケーション開発者は、この情報を使用して、作業を行うクラスタ環境を理解できます。

次の図は、クラスタ管理の概念がクラスタの構造にどのように対応するかを示しています。

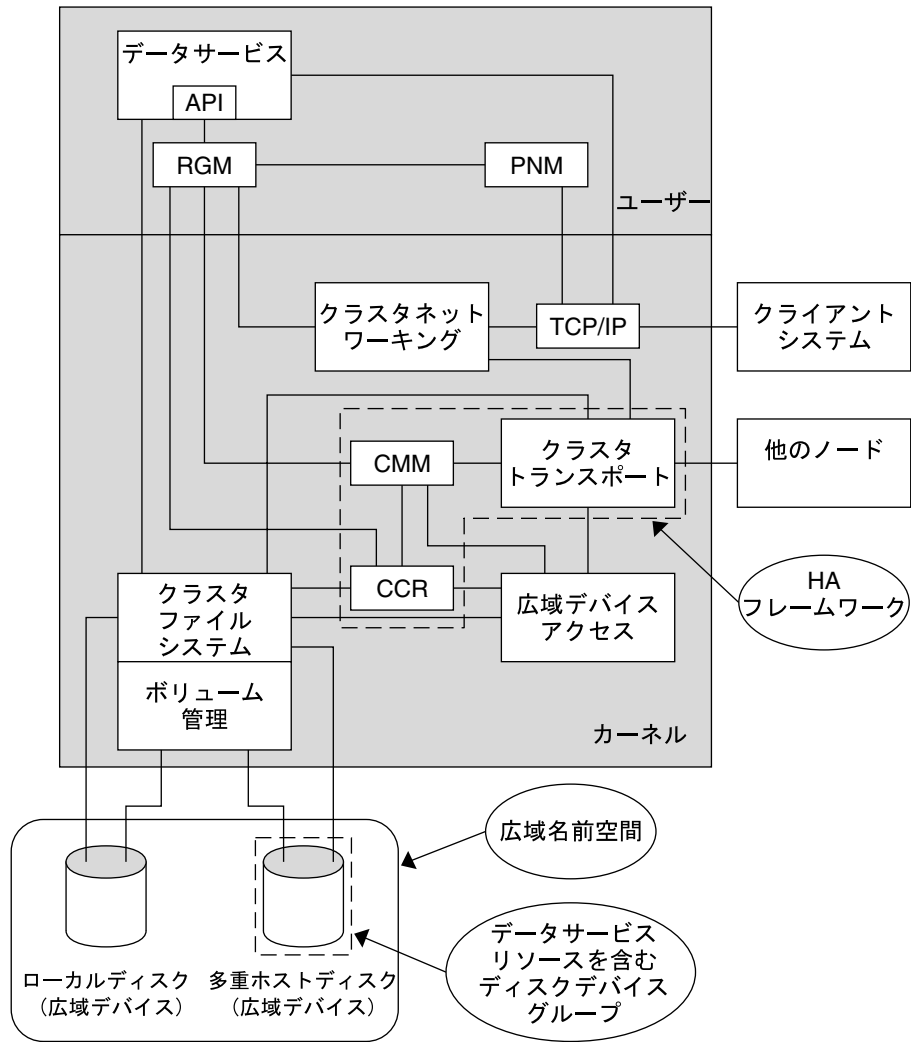


図 3-1 Sun Cluster ソフトウェアの構造

## 管理インタフェース

任意のユーザーインタフェースを使用し、Sun Cluster および Sun Cluster データサービスをインストール、構成、管理できます。システム管理作業は、コマンド行インタフェースで行えます。コマンド行インタフェースには、構成作業を簡略化するユーティリティがあります。Sun Cluster には、Sun Management Center の一部として実行される、特定のクラスタ作業に GUI を提供するモジュールもあります。

管理インタフェースの詳細については、『*Sun Cluster 3.0* のシステム管理』の概要を説明する章を参照してください。

## クラスタ内の時間

クラスタ内のすべてのノード間の時刻は同期をとる必要があります。クラスタノードの時刻と外部の時刻ソースの同期をとるかどうかは、クラスタの操作にとって重要ではありません。Sun Cluster は、Network Time Protocol (NTP) を使用して、ノード間のクロックの同期をとっています。

通常、システムクロックが数分の 1 秒程度変更されても問題は起こりません。しかし、システムクロックと時刻の起点の同期をとるため

に、`date(1)`、`rdate(1M)`、`xntpdate(1M)` を (対話形式または cron スクリプト内で) アクティブクラスタに対して実行すると、これよりも大幅に時刻変更が可能です。この強制的な変更は、ファイル修正時刻の表示に問題が生じたり、NTP サービスに混乱が生じる場合があります。

Solaris オペレーティング環境を各クラスタノードにインストールする場合は、ノードのデフォルトの時刻と日付の設定を変更できます。通常は、工場出荷時のデフォルト値を使用します。

`scinstall(1M)` を使用して Sun Cluster をインストールする場合、インストールプロセスの手順の 1 つとして、クラスタの NTP を構成します。Sun Cluster は、テンプレートファイル `ntp.cluster` を提供します (インストールされたクラスタノードの `/etc/inet/ntp.cluster` を参照)。これは、1 つのノードを優先ノードにし、すべてのクラスタノード間で対等関係を確立します。ノードは、そのプライベートホスト名によって識別され、時刻の同期化はクラスタインターCONNECT全体で発生します。NTP のクラスタを構成する方法については、『*Sun Cluster 3.0* ソフトウェアのインストール』を参照してください。

また、クラスタの外部に 1 つまたは複数の NTP サーバーを設定し、`ntp.conf` ファイルを変更してその構成を反映させることもできます。

通常の操作では、クラスタの時刻を調整する必要はありません。ただし Solaris オペレーティング環境をインストールしたときに時刻が正しく設定されておらず、変更する必要がある場合は、『*Sun Cluster 3.0* ソフトウェアのインストール』を参照してください。



## 高可用性フレームワーク

Sun Cluster では、ネットワークインタフェース、アプリケーションそのもの、ファイルシステム、および多重ホストディスクなど、ユーザーとデータ間のパスにおけるすべてのコンポーネントの可用性が高くなっています。一般に、クラスターコンポーネントは、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稼働し続けられる場合は可用性が高くなります。

次の表は、Sun Cluster コンポーネントの障害の種類 (ハードウェアとソフトウェアの両方) と、高可用性フレームワークに組み込まれた回復の種類を示しています。

表 3-1 Sun Cluster の障害の検出と回復のレベル

障害が発生した クラスタリソース	ソフトウェアの回復	ハードウェアの回復
データサービス	HA API、HA フレームワーク	なし
パブリックネットワークアダプタ	ネットワークアダプタのフェイルオーバー (NAFO)	複数のパブリックネットワークアダプタカード
クラスタファイルシステム	主複製と二次複製	多重ホストディスク
ミラー化された多重ホストディスク	ボリューム管理 (Solstice DiskSuite および VERITAS Volume Manager)	ハードウェア RAID-5 (Sun StorEdge A3x00 など)
広域デバイス	主複製と二次複製	デバイス、クラスタトランスポート接続点への多重パス
プライベートネットワーク	HA トランスポートソフトウェア	ハードウェアから独立した多重プライベートネットワーク
ノード	CMM、failfast ドライバ	複数ノード

Sun Cluster の高可用性フレームワークは、ノードの障害を素早く検出して、クラスター内の残りのノードにあるフレームワークリソース用に新しい同等のサーバーを作成します。すべてのフレームワークリソースが使用できなくなることはありません。障害が発生したノードによって影響を受けないフレームワークリソースは、回復中も完全に使用できます。さらに、障害が発生したノードのフレームワークリ

ソースは、回復されると同時に使用可能になります。回復されたフレームワークリソースは、他のすべてのフレームワークリソースが回復するまで待機する必要はありません。

最も可用性の高いフレームワークリソースは、リソースを使用してアプリケーション(データサービス)に透過的に回復されます。フレームワークリソースのアクセス方式は、ノードの障害時にも完全に維持されます。アプリケーションは、フレームワークリソースサーバーが別のノードに移動したことを認識できないだけです。単一ノードの障害は、別ノードからのディスクに対する代替ハードウェアパスが存在するかぎり、ファイル、デバイス、およびディスクボリュームを使用する残りのノード上のプログラムに対して完全に透過的です。この例としては、複数ノードへのポートを持つ多重ホストディスクの使用があります。

## クラスタメンバーシップモニター

クラスタメンバーシップモニター (CMM) は、エージェントの分散型セットであり、クラスタメンバーごとに1つあります。これらのエージェントは、クラスタインターコネクトによってメッセージを交換して、次のことを行います。

- すべてのノード(定足数/quorum)で一貫したメンバーシップの表示を行う
- メンバーシップの変更に応じて、登録されたコールバックを使用して同期化された再構成を行う
- クラスタパーティション分割を処理する (split-brain と amnesia)
- すべてのクラスタメンバー間での完全な接続を保証する

Sun Cluster の前のリリースとは異なり、CMM は完全にカーネルで実行されます。

### クラスタメンバーシップ

CMM の主な機能は、特定の時刻にクラスタに属するノードの集合に対して、クラスタ全体の同意を確立することです。Sun Cluster は、この制約をクラスタメンバーシップと呼びます。

クラスタメンバーシップを決定して、最終的にデータの完全性を保証するために、CMM は次のことを行います。

- クラスタへのノードの結合、またはクラスタからのノードの切り離しなど、クラスタメンバーシップの変更を考慮する
- 障害のあるノードがクラスタから切り離されるように保証する

- 障害のあるノードが、修復されるまでクラスタの外部におかれるように保証する
- 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する

クラスタが複数の独立したクラスタに分割されないように防止する方法については、52ページの「定足数と定足数デバイス」を参照してください。

### クラスタメンバーシップモニターの再構成

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 (アプリケーション) 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、クラスタの同期化構成を実行します。これにより、クラスタリソースは、クラスタの新しいメンバーシップに基づいて再分配されます。

### クラスタ構成レポジトリ (CCR)

クラスタ構成レポジトリ (CCR) は、クラスタの構成と状態に関する情報を保存するための、プライベートなクラスタ全体のデータベースです。CCR は分散データベースです。各ノードは、このデータベースの完全なコピーを維持します。CCR は、すべてのノードで、クラスタ全体の一貫した表示が行われるように保証します。データの破壊を避けるために、各ノードは、クラスタリソースの現在の状態を知る必要があります。

CCR は、可用性の高いサービスとしてカーネルに実装されます。

CCR は、更新に 2 フェーズのコミットアルゴリズムを使用します。更新はすべてのクラスタメンバーで正常に終了しなければなりません。そうしないと、その更新はロールバックされます。CCR はクラスタインターコネクトを使用して、分散更新を適用します。



---

**注意** - CCR はテキストファイルから構成されますが、CCR ファイルは決して手作業で編集しないでください。各ファイルには、一貫性を保証するための検査合計レコードが含まれます。CCR ファイルを手作業で更新すると、ノードまたはクラスタ全体の機能が停止する可能性があります。

---

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。

## 広域デバイス

Sun Cluster は、広域デバイスを使用して、デバイスが物理的に接続されている場所に関係なく、任意のノードからクラスタ内のすべてのデバイスに対して、クラスタ全体の可用性の高いアクセスを可能にします。通常、広域デバイスへのアクセスを提供しているときにノードに障害が発生すると、Sun Cluster はそのデバイスへの別のバスを自動的に検出して、そのバスにアクセスを切り替えます。Sun Cluster 広域デバイスには、ディスク、CD-ROM、テープが含まれます。ディスクは、唯一サポートされている多重ポート広域デバイスです。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 を自動的に割り当てます。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。広域デバイス名前空間は、`/dev/global` ディレクトリにあります。詳細については、47ページの「広域名前空間」を参照してください。

多重ポート広域デバイスは、1つのデバイスに対して複数のバスを提供します。多重ホストディスクの場合、ディスクは複数のノードがホストとなるディスクデバイスグループの一部であるため、多重ホストディスクの可用性は高くなります。

## デバイス ID (DID)

Sun Cluster は、デバイス ID (DID) 擬似ドライバと呼ばれる構造によって広域デバイスを管理します。このドライバは、多重ホストディスク、テープドライブ、CD-ROM を含むクラスタ内のすべてのデバイスに対して、一意の ID を自動的に割り当てるために使用されます。

デバイス ID (DID) 擬似ドライバは、クラスタの広域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探索して、一意のディスクデバイスのリストを作成し、それぞれに対して、クラスタのすべてのノードで一貫した一意のメジャー番号およびマイナー番号を割り当てます。広域デバイスへのアクセスは、ディスクを示す `c0t0d0` などの従来の Solaris デバイス ID ではなく、DID ドライバによって割り当てられた一意のデバイス ID を利用して行われます。

この方法により、ディスクデバイスを利用するすべてのアプリケーション (ボリューム管理ソフトウェアまたは raw デバイスを使用するアプリケーション) が、一貫したパスを使用してデバイスにアクセスできます。各デバイスのローカルメジャー番号およびマイナー番号はノードによって異なり、Solaris デバイス命名規則も変更する可能性があるため、この一貫性は、多重ホストディスクにとって特に重要です。たとえば、ノード 1 は多重ホストディスクを `c1t2d0` と表示し、同じディスクをノード 2 は `c3t2d0` と表示する場合があります。DID ドライバは、そのノードが代わりに使用する `d10` などの広域名を割り当てて、各ノードに対して多重ホストディスクとの一貫したマッピングを与えます。

デバイス ID は、`scdidadm(1M)` および `scgdevs(1M)` によって更新、管理します。詳細については、それぞれのマニュアルページを参照してください。

## ディスクデバイスグループ

Sun Cluster では、すべての多重ホストディスクが、Sun Cluster フレームワークの制御のもとになければなりません。まず、ボリューム管理ソフトウェアのディスクグループ (Solstice DiskSuite ディスクセットか、VERITAS Volume Manager ディスクグループのどちらか) を多重ホストディスクに作成します。次に、ボリューム管理ソフトウェアのディスクグループを Sun Cluster のディスクデバイスグループとして登録します。ディスクデバイスグループは、広域デバイスの一種です。さらに、Sun Cluster は、個々のディスクすべてをディスクデバイスグループとして割り当てます。

---

注 - ディスクデバイスグループは、リソースグループとは別のものです。あるノードが1つのリソースグループ(データサービスプロセスのグループを表す)をマスターする一方で、別のノードが、データサービスによってアクセスされるディスクグループをマスターできます。ただし、最も良い方法は、特定のアプリケーションのデータを保存するディスクデバイスグループと、アプリケーションのリソース(アプリケーションデーモン)を同じノードに含むリソースグループを維持することです。ディスクデバイスグループとリソースグループの関連付けの詳細については、『Sun Cluster 3.0 データサービスのインストールと構成』の概要を示す章を参照してください。

---

ディスクデバイスグループでは、ボリューム管理ソフトウェアのディスクグループは実際に使用するディスクに対してマルチパスサポートを提供するため、広域になります。多重ホストディスクに物理的に接続された各クラスタノードは、ディスクデバイスグループへのパスを提供します。

---

注 - 広域デバイスは、複数のクラスタノードがホストとなるデバイスグループの一部である場合、可用性が高くなります。

---

## ディスクデバイスフェイルオーバー

ディスク格納装置は複数のノードに接続されるため、現在デバイスグループをマスターしているノードに障害が生じた場合、その格納装置にあるすべてのディスクデバイスグループには、代替パスによってアクセスできます。デバイスグループをマスターするノードの障害は、回復と一貫性の検査を実行するために要する時間を除けば、デバイスグループへのアクセスに影響しません。この時間中、デバイスグループが使用可能になるまで、すべての要求は(アプリケーションには透過的に)阻止されます。

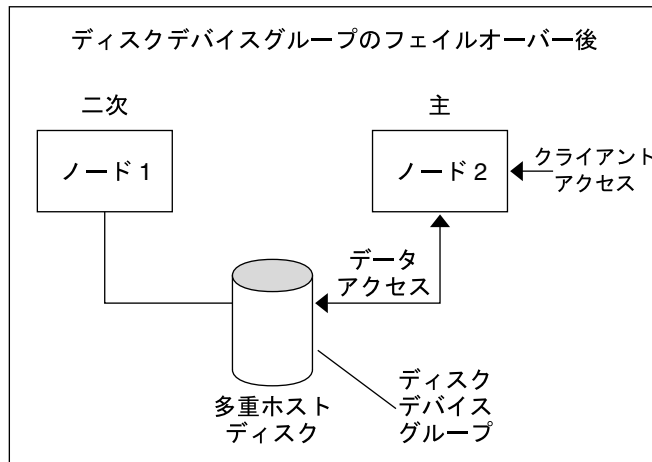
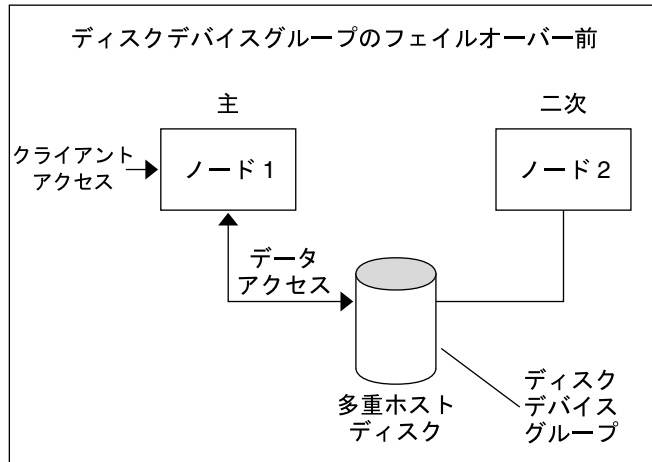


図 3-2 ディスクデバイスグループのフェイルオーバー

## 広域名前空間

広域デバイスを有効にする Sun Cluster の機構は、広域名前空間です。広域名前空間には、ボリューム管理ソフトウェアの名前空間とともに、`/dev/global/` 階層が含まれます。広域名前空間は、多重ホストディスクとローカルディスクの両方（および CD-ROM やテープなどの他のクラスタデバイスすべて）を反映して、多重ホストディスクへの複数のフェイルオーバーパスを提供します。多重ホストディスクに物理的に接続された各ノードは、クラスタ内のすべてのノードの記憶装置に対するパスを提供します。

通常、ボリューム管理ソフトウェアの名前空間は、Solstice DiskSuite の場合、`/dev/md/diskset/dsk` および `/dev/md/diskset/rdisk` ディレクトリにあります。また、VxVM の場合は、`/dev/vx/dsk/disk-group` および `/dev/vx/rdisk/disk-group` の各ディレクトリにあります。これらの名前空間は、それぞれクラスタを介してインポートされた各 Solstice DiskSuite ディスクセットと各 VxVM ディスクグループのディレクトリからなります。これらの各ディレクトリには、そのディスクセットまたはディスクグループ内の各メタデバイスまたはボリュームのデバイスノードが入っています。

Sun Cluster では、ローカルのボリューム管理ソフトウェアの名前空間の各デバイスノードが、`/global/.devices/node@nodeID` ファイルシステム内のデバイスノードに対するシンボリックリンクによって置き換えられています。ここで、*nodeID* は、クラスタ内のノードを表す整数になります。Sun Cluster は、その標準的な場所に引き続きシンボリックリンクとしてボリューム管理ソフトウェアデバイスも表示します。広域名前空間と標準ボリューム管理ソフトウェア名前空間の両方を任意のクラスタノードから使用できます。

広域名前空間には、次の利点があります。

- 各ノードの独立性が高く、デバイス管理モデルをほとんど変更しなくてすむ。
- デバイスを選択的に広域にできる。
- Sun の製品以外のリンクジェネレータが引き続き動作する。
- ローカルデバイス名を指定すると、その広域名を取得するために簡単なマッピングが提供される。

## ローカル名前空間と広域名前空間の例

次の表は、多重ホストディスク `c0t0d0s0` でのローカル名前空間と広域名前空間のマッピングを示しています。

表 3-2 ローカル名前空間と広域名前空間のマッピング

コンポーネント パス	ローカルノード 名前空間	広域名前空間
Solaris 論理名	<code>/dev/dsk/ c0t0d0s0</code>	<code>/global/.devices/node@ID/dev/dsk/c0t0d0s0</code>
DID 名	<code>/dev/did/ dsk/d0s0</code>	<code>/global/.devices/node@ID/dev/did/dsk/d0s0</code>



表 3-2 ローカル名前空間と広域名前空間のマッピング 続く

コンポーネント /パス	ローカルノード 名前空間	広域名前空間
Solstice DiskSuite	/dev/md/ diskset/dsk/d0	/global/.devices/node@ID/dev/md/diskset/ dsk/d0
VERITAS Volume Manager	/dev/vx/dsk/ disk-group/v0	/global/.devices/node@ID/dev/vx/dsk/ disk-group/v0

広域名前空間はインストール時に自動的に生成されて、再構成再起動のたびに更新されます。広域名前空間は、`scgdevs (1M)` コマンドを実行することでも生成できます。

## 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、1つのノード上のカーネル、配下のファイルシステムおよびディスクへの物理接続を持つノードで実行されるボリューム管理ソフトウェアとの間のプロキシです。

クラスタファイルシステムは、1つまたは複数のノードへの物理接続を持つ広域デバイス (ディスク、テープ、CD-ROM) に依存しています。広域デバイスは、ノードが記憶装置への物理接続を持つかどうかに関係なく、同じファイル名 (たとえば、`/dev/global/`) によってクラスタ内のすべてのノードからアクセスできます。広域デバイスは通常のデバイスと同様に使用できます。つまり、`newfs` または `mkfs`、あるいはこの両方を使用し、ファイルシステムを作成できます。

広域デバイスには、`mount -g` を使用して広域に、または `mount` を使用してローカルにファイルシステムをマウントできます。プログラムは、同じファイル名 (たとえば、`/global/foo`) によって、クラスタ内のすべてのノードからクラスタファイルシステムのファイルにアクセスできます。クラスタファイルシステムは、すべてのクラスタメンバーにマウントされます。クラスタファイルシステムをクラスタメンバーのサブセットにマウントすることはできません。

## クラスタファイルシステムの使用法

Sun Cluster では、すべての多重ホストディスクがディスクデバイスグループとして構成されています。これは、Solstice DiskSuite ディスクセット、VxVM ディスクグループ、またはソフトウェアベースのボリューム管理ソフトウェアの制御下でない個々のディスクが該当します。また、ローカルディスクも、パスが各ノードから各ローカルディスクにつながるという点で、ディスクデバイスグループとして構成されています。この設定では、すべてのノードからディスク上のデータを利用できるとは限りません。データは、ディスク上のファイルシステムがクラスタファイルシステムとして広域にマウントされた場合にのみ、すべてのノードに利用可能になります。

クラスタファイルシステムに組み込まれたローカルファイルシステムには、ディスク記憶装置への単一の接続だけがあります。ディスク記憶装置への物理接続を持つノードに障害が生じると、他のノードからクラスタファイルシステムへのアクセスが不可能になります。また、他のノードから直接アクセスできない単一ノードにローカルファイルシステムを置くこともできます。

HA データサービスは、サービス用のデータが、クラスタファイルシステム内のディスクデバイスグループに保存されるように設定されています。この設定にはいくつかの利点があります。まず、データの可用性が高くなります。つまり、ディスクは多重ホスト化されているため、現在の主ノードからのパスに障害が発生した場合、アクセスは同じディスクへの直接アクセスを持つ別のノードに切り替えられます。次に、データはクラスタファイルシステムにあるため、任意のクラスタノードから直接表示できます。現在ディスクデバイスグループをマスターしているノードにログインしてデータを表示する必要はありません。

## プロキシファイルシステム (PXFS)

クラスタファイルシステムは、プロキシファイルシステム (PXFS) に基づいています。これには、次の機能があります。

- PXFS を使用すると、ファイルのアクセス場所が透過的になります。プロセスはファイルがシステム内のどこに置かれていても開くことができ、また、すべてのノード上のプロセスが同じパス名を使用してファイルを見つけられます。
- PXFS は、一貫したプロトコルを使用して、ファイルが複数のノードから同時にアクセスされている場合でも、UNIX ファイルアクセス方式を維持します。
- PXFS には、拡張キャッシュ機能と、大量のデータオブジェクトを効率的に移動するためのゼロコピーバルク入出力移動機能があります。

- PXFS を使用すると、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る限り、障害を検出しません。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。
- PXFS は、実際のファイルシステムおよびボリューム管理ソフトウェアに依存していません。PXFS 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。
- PXFS は、既存の Solaris ファイルシステムの最上位の vnode インタフェースに構築されています。このインタフェースを使用すると、拡張カーネル修正なしで、PXFS を実装できます。

PXFS は、明確なファイルシステム形式ではありません。つまり、クライアントには、実際のファイルシステム (たとえば UFS) が表示されます。

## クラスタファイルシステムの独立性

クラスタファイルシステムは、実際のファイルシステムおよびボリューム管理ソフトウェアに依存していません。現在、クラスタファイルシステムは、Solstice DiskSuite または VERITAS Volume Manager のどちらかを使用して、UFS で作成できます。

通常のファイルシステムと同様、クラスタファイルシステムは 2 つの方法でマウントできます。

- 手作業によるマウント— mount コマンドと -g オプションを使用して、コマンド行からクラスタファイルシステムをマウントします。次に例を示します。

```
# mount -g /dev/global/dsk/d0s0 /global/oracle/data
```

- 自動マウント— global マウントオプションによって /etc/vfstab ファイルにエントリを作成します。さらに、すべてのノードの /global ディレクトリ下にマウントポイントを作成します。ディレクトリ /global を推奨しますが、他の場所でも構いません。次に、/etc/vfstab ファイルの、クラスタファイルシステムを示す行の例を示します。

```
/dev/md/oracle/dsk/d1 /dev/md/oracle/rdisk/d1 /global/oracle/  
data ufs 2 yes global,logging
```

---

注 - Sun Cluster には、クラスタファイルシステムの命名ポリシーはありませんが、`/global/disk-device-group` など、同じディレクトリのもとにすべてのクラスタファイルシステムのマウントポイントを作成すると、管理が簡単になります。詳細については、『Sun Cluster 3.0 ソフトウェアのインストール』および『Sun Cluster 3.0 のシステム管理』を参照してください。

---

### **syncdir** マウントオプション

クラスタファイルシステムには、`syncdir` マウントオプションを使用できますが、`syncdir` を指定しない方がパフォーマンスは向上します。`syncdir` を指定すると、POSIX 準拠の書き込みが保証されます。指定しないと、UFS ファイルシステムの場合と同じ動作となります。たとえば、`syncdir` を指定しないと、場合によっては、ファイルを閉じるまでスペース不足条件を検出できません。`syncdir` (および POSIX 動作) を指定すると、スペース不足条件は書き込み動作中に検出されます。`syncdir` を指定しないことで生じる問題はほとんどないため、このオプションを指定しないで、パフォーマンスを向上させることを推奨します。

広域デバイスとクラスタファイルシステムについては、74ページの「ファイルシステムに関する FAQ」を参照してください。

## 定足数と定足数デバイス

クラスタノードはデータとリソースを共有するため、クラスタはデータとリソースの完全性を維持するための手順に従う必要があります。あるノードがメンバーシップに関するクラスタの規則に合致しない場合、クラスタは、そのノードがクラスタに属するのを禁止する必要があります。

Sun Cluster では、クラスタへのノードの結合を判断する機構を定足数 (*quorum*) と呼びます。Sun Cluster は、多数決のアルゴリズムを使用して定足数を実装します。クラスタノードと定足数デバイスは、どちらも複数のノード間で共有されるディスクであり、定足数を確立するために投票 (*vote*) します。定足数デバイスにはユーザーデータを含むことができます。

定足数アルゴリズムは動的に動作します。クラスタイベントによってその計算が発生した場合、計算結果はクラスタの存続期間中、変化し続けます。定足数は、`split-brain` と `amnesia` という2つのクラスタの問題を防止します。どちらも、一貫

性のないデータがクライアントに提示される原因となる可能性があります。次の表は、これらの2つの問題と、定足数によるその解決方法を説明しています。

表 3-3 クラスタ定足数、および split-brain と amnesia の問題

問題	説明	定足数による解決策
split brain	ノード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分割され、それぞれが唯一のパーティションであると誤って認識した場合に発生する	過半数の投票があるパーティション (サブクラスタ) だけが、クラスタとして実行できるようにする (1つのパーティションだけが、このような過半数によって存在できる)
amnesia	停止時よりも古いクラスタデータを使用して、停止後にクラスタが再起動すると発生する	クラスタの起動時に、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、最新の構成を持つ) 少なくとも1つのノードが存在するように保証する

## 定足数投票数

クラスタノードと定足数デバイス (複数のノード間で共有されるディスク) 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。またノードは、たとえば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ときには、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、投票数がゼロ以外で、定足数デバイスへのポートを持つノードの数を示します。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。

定足数デバイスは、クラスタのインストール中、または『Sun Cluster 3.0 のシステム管理』で説明している手順を使用して後で構成します。

---

注 - 定足数デバイスは、現在接続されている少なくとも1つのノードがクラスタメンバーである場合にのみ、投票数を獲得します。また、クラスタの起動中、定足数デバイスは、現在接続されている少なくとも1つのノードが起動中で、その停止時に最も最近起動されたクラスタのメンバーであった場合にのみ投票数を獲得します。

---

## 定足数の構成

定足数 (quorum) の構成は、クラスタ内のノードの数によって異なります。

- **2 ノードクラスタ** - 2 ノードクラスタを形成するには、定足数投票数 (quorum vote count) が2つ必要です。これらの2つの投票数は、2つのクラスタノード、または1つのノードと1つの定足数デバイスのどちらかによるものです。ただし、2 ノードクラスタでは、一方のノードに障害が発生したときに単一ノードで処理を続行できるように、1つの定足数デバイスが構成されなければなりません。
- **3 ノード以上のクラスタ** - ディスク格納装置へのアクセスを共有するすべてのノードペア間に、定足数デバイスを指定する必要があります。たとえば、図3-3のような3 ノードクラスタを想定します。この場合、ノードAとノードBが同じディスク格納装置へのアクセスを共有し、ノードBとノードCは別のディスク格納装置へのアクセスを共有します。この構成では、合計5つの投票数があります。3つはノードによるもの、2つはノード間で共有される定足数デバイスによるものです。クラスタを形成するには、過半数の投票数(この場合は3つ)を必要とします。定足数デバイスを、ディスク格納装置へのアクセスを共有するすべてのノードペア間に指定することは、Sun Cluster では必要とされておらず、また実行もされません。しかし、N+1 構成が2 ノード構成になり、両方のディスク格納装置へアクセスするノードに障害が発生すると、必要な投票数を提供しません。すべてのペア間で定足数デバイスを構成すると、残りのノードはクラスタとして動作を継続することができます。

これらの構成の例については、図3-3を参照してください。

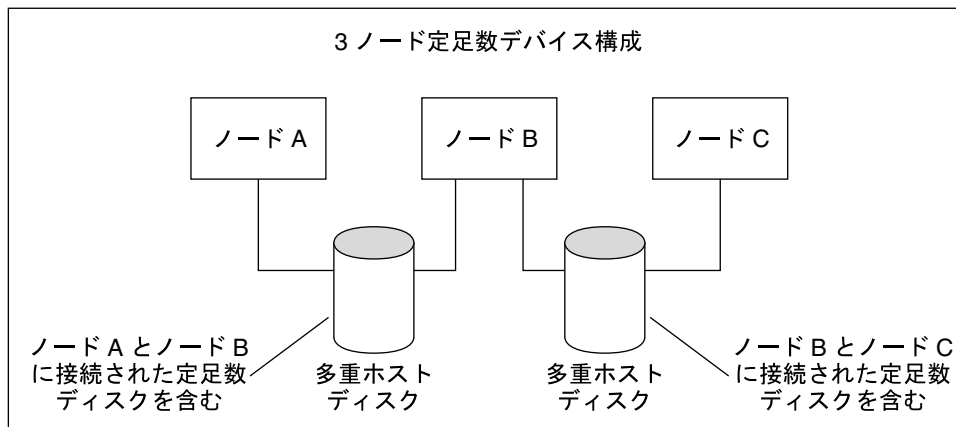
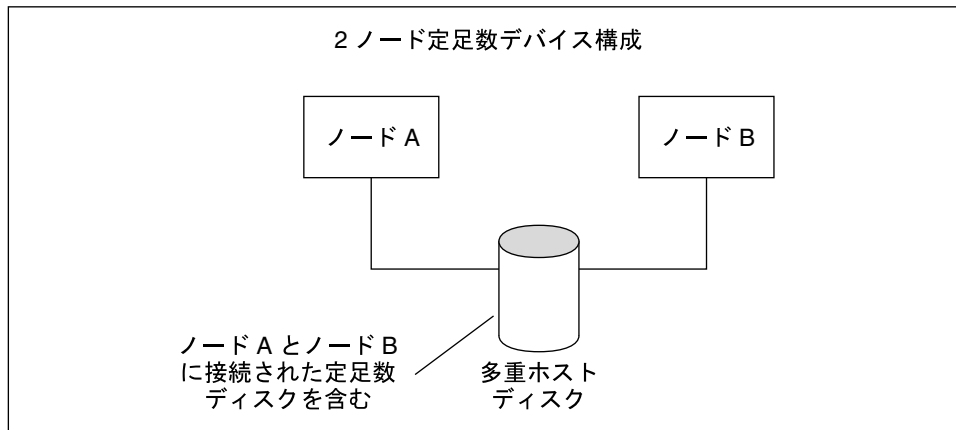


図 3-3 定足数デバイス構成の例

## 定足数のガイドライン

定足数デバイスを設定するときは、次のガイドラインを使用してください。

- 同じ共有ディスク格納装置に接続されたすべてのノード間で定足数デバイスを確立します。共有格納装置内に 1 つのディスクを定足数デバイスとして追加し、いずれかのノードに障害が生じたときに、他のノードが定足数を維持して、共有格納装置上のディスクデバイスグループをマスターできるようにします。
- 定足数デバイスは少なくとも 2 つのノードに接続する必要があります。
- 定足数デバイスとして、二重ポート定足数デバイスとして使用される、任意の SCSI-2 または SCSI-3 ディスクが使用できます。3 つ以上のノードに接続されたディスクは、ディスクが定足数デバイスとして使用されるかどうかに関係なく、

SCSI-3 持続的グループ予約 (PGR) をサポートしなければなりません。詳細については、『Sun Cluster 3.0 ソフトウェアのインストール』の計画に関する章を参照してください。

- 定足数デバイスとしてユーザーデータを含むディスクを使用できます。

---

ヒント・ノードの集合間で複数の定足数デバイスを構成してください。異なる格納装置から複数のディスクを使用して、奇数の定足数デバイスをノードの各集合間で構成してください。これにより、個々の定足数デバイスの障害から保護できます。

---

## 障害による影響の防止

クラスタの主要な問題は、クラスタがパーティション分割される (sprit-brain と呼ばれる) 原因となる障害です。この障害が発生すると、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ノードの一部が、個々のクラスタまたはクラスタの一部を形成しようとしています。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。複数のノードがディスクに書き込もうとすると、データが破壊される可能性があります。

障害による影響の防止機能では、多重ホストディスクへのアクセスを物理的に防止することによって制限します。障害が発生するかパーティション分割され、ノードがクラスタから切り離されると、障害による影響の防止機能によって、ノードがディスクにアクセスできなくなります。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。

ディスクデバイスサービスは、多重ホストディスクを使用するサービスに対して、フェイルオーバー機能を提供します。現在、ディスクデバイスグループの主ノード (所有者) として機能しているクラスタメンバーに障害が発生するか、またはこのメンバーに到達できなくなると、新しい主ノードが選択されて、ディスクデバイスグループへのアクセスが可能になり、わずかな割り込みだけで処理が続行されます。このプロセス中、古い主ノードは、新しい主ノードが起動される前に、デバイスへのアクセスを放棄しなければなりません。ただし、あるメンバーがクラスタから切り離されて到達不能になると、クラスタはそのノードに対して、主ノードであったデバイスを解放するように通知できません。したがって、存続するメンバーが、障害の発生したメンバーから広域デバイスを制御してアクセスできるようにする手段が必要です。

Sun Cluster は、SCSI ディスク予約を使用して、障害による影響の防止機能を実装します。SCSI 予約を使用すると、障害が発生したノードは、多重ホストディスクによって阻止されて、これらのディスクへのアクセスが防止されます。



SCSI-2 ディスク予約は、ある形式の予約をサポートしています。これは、ディスクに接続されたすべてのノードへのアクセスを付与するか (予約が設定されていない場合)、または単一ノード (予約を保持するノード) へのアクセスを制限するものです。

クラスタメンバーは、別のノードがクラスタインターコネクトを介して通信していないことを検出すると、障害による影響の防止手順を開始して、そのノードが共有ディスクへアクセスするのを防止します。この障害による影響の防止機能が実行される場合、通常、阻止されるノードは、そのコンソールに「reservation conflict」(予約の衝突) というメッセージを表示して停止します。

予約の衝突は、ノードがクラスタメンバーではなくなったことが検出された後で、SCSI 予約がこのノードと他のノードの間で共有されるすべてのディスクに対して設定されると発生します。阻止されるノードは阻止されていることを認識しない場合があります、共有ディスクのどれかにアクセスしようとして、予約を検出して停止します。

## ボリューム管理ソフトウェア

Sun Cluster は、ボリューム管理ソフトウェアを使用して、ミラーとホットスペアディスクによりデータの可用性を向上させ、ディスクの障害と交換を処理します。

Sun Cluster には、独自の内部ボリューム管理ソフトウェアコンポーネントはありませんが、次のボリューム管理ソフトウェアに依存しています。

- Solstice DiskSuite
- VERITAS Volume Manager

クラスタ内のボリューム管理ソフトウェアは、次のものに対するサポートを提供しています。

- ノード障害のフェイルオーバー処理
- 異なるノードからの多重パスサポート
- ディスクデバイスグループへの遠隔からの透過アクセス

Sun Cluster でボリューム管理ソフトウェアを設定する場合は、ボリューム管理ソフトウェアのディスクグループのラッパーである Sun Cluster ディスクデバイスとして、多重ホストディスクを構成します。このデバイスは、Solstice DiskSuite ディスクセットまたは VxVM ディスクグループのどちらかにできます。

データサービスに使用されるディスクグループをミラー化用に構成して、クラスタ内でのディスクの可用性を高める必要があります。

メタデバイスまたはプレックスを、raw デバイス (データベースアプリケーション) として、または UFS ファイルシステムの保持用に使用できます。

ボリューム管理オブジェクトのメタデバイスとボリュームは、クラスタの制御下に置かれるため、ディスクデバイスグループになります。たとえば、Solstice DiskSuite では、(metaset (1M) コマンドを使用して) クラスタにディスクセットを作成すると、同じ名前の対応するディスクデバイスグループが作成されます。さらに、そのディスクセットにメタデバイスを作成すると、それらは広域デバイスになります。したがって、ディスクセットはディスクデバイス (DID デバイス) と、セット内のデバイスすべてのポート先であるホストの集合になります。クラスタ内のすべてのディスクセットを、セット内の複数のホストによって作成して、HA を実現する必要があります。同様の状態は、VERITAS Volume Manager を使用した場合にも発生します。各ボリューム管理ソフトウェアの設定の詳細については、『Sun Cluster 3.0 ソフトウェアのインストール』の付録を参照してください。

ディスクセットまたはディスクグループを計画する場合の重要な考慮事項として、その関連ディスクデバイスグループが、クラスタ内のアプリケーションリソース (データ) にどのように関連付けられているかを理解する必要があります。詳細は、『Sun Cluster 3.0 ソフトウェアのインストール』と『Sun Cluster 3.0 データサービスのインストールと構成』を参照してください。

## データサービス

データサービスという用語は、単一サーバーではなくクラスタでの実行用に構成されたアプリケーション (Apache Web Server など) を表します。データサービスには、アプリケーションソフトウェアと、そのアプリケーションの起動、停止、監視を行う Sun Cluster ソフトウェアが含まれます。

Sun Cluster には、クラスタ内のアプリケーションを制御して監視するために使用されるデータサービス方式があります。これらの方式は、クラスタノード上のアプリケーションの起動、停止、および監視にこれらの方式を使用する、リソースグループマネージャ (RGM) の制御のもとで実行されます。これらの方式を、クラスタフレームワークソフトウェアおよび多重ホストディスクとともに使用すると、アプリケーションは可用性の高いデータサービスになります。可用性の高いデータサービスとなったアプリケーションは、クラスタ内で単一の障害が発生した後で起こる深刻なアプリケーション割り込みを防止できます。障害は、ノード、インタ

フェースコンポーネント、またはアプリケーションそのものに対して発生する可能性があります。

RGM は、アプリケーションとネットワークリソース (論理ホスト名と共有アドレス) のインスタンスを含め、クラスタ内のリソースも管理します。

Sun Cluster には、アプリケーションプログラマが、他のアプリケーションを可用性の高いデータサービスとして Sun Cluster で実行するのに必要なデータサービス方式を開発するための API とデータサービス開発ツールもあります。

## リソースグループマネージャー (RGM)

Sun Cluster は、アプリケーションの可用性を高めるか、またはスケーラブルにするための環境を提供します。RGM は、リソースに対して作用します。リソースは、次のいずれかを行える論理コンポーネントです。

- オンラインおよびオフラインにできる (切り替えられる)
- RGM フレームワークによって管理できる
- 単一ノード (フェイルオーバーモード) または複数ノード (スケーラブルモード) でホストできる

RGM は、データサービス (アプリケーション) を、リソースタイプの実装によって管理される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、汎用データサービステンプレート、DSDL API (データサービス開発ライブラリ API)、Sun Cluster RMAPI (リソース管理 API) のどれかを使用して作成されます。クラスタ管理者は、リソースグループと呼ばれるコンテナにリソースを作成して管理します。これは、フェイルオーバーとスイッチオーバーの基本ユニットになります。RGM は、クラスタメンバーシップの変更に応じて、指定ノードのリソースグループを停止および開始します。

## フェイルオーバーデータサービス

データサービスが実行されているノード (主ノード) に障害が発生すると、サービスは、ユーザーによる介入なしで別の作業ノードに移行します。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利用します。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。

フェイルオーバーデータサービスでは、アプリケーションインスタンスは単一ノードでのみ実行されます。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) しようとしています。

## スケーラブルデータサービス

スケーラブルデータサービスは、複数ノードのアクティブインスタンスに対して効果があります。スケーラブルサービスは、スケーラブルリソースグループを使用して、アプリケーションリソースを含め、フェイルオーバーリソースグループを使用して、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) を含めます。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一度に実行できます。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のホストとなるすべてのノードは、同じ共有アドレスを使用して、サービスに対するホストとなります。

サービス要求は、単一ネットワークインタフェース (広域インタフェース、または GIF) を介してクラスタに入り、負荷均衡ポリシーによって設定されたいくつかの定義済みアルゴリズムの1つに基づいてノードに分配されます。クラスタは、負荷均衡ポリシーを使用して、いくつかのノード間でサービス負荷の均衡をとることができます。他の共有アドレスをホストしている別のノード上に、複数の GIF が存在する可能性があります。

スケーラブルサービスの場合、アプリケーションインスタンスはいくつかのノードで同時に実行されます。広域インタフェースのホストとなるノードに障害が発生すると、広域インタフェースは別のノードで処理を続行します。アプリケーションインスタンスの実行に失敗した場合、そのインスタンスは同じノードで再起動しようとしています。

アプリケーションインスタンスを同じノードで再起動できず、別の未使用のノードがサービスを実行するように構成されている場合、サービスはその未使用ノードで処理を続行します。あるいは、残りのノードで実行し続けて、サービススループットを低下させることとなります。

---

注 - 各アプリケーションインスタンスの TCP 状態は、GIF ノードではなく、インスタンスを持つノードで維持されます。したがって、GIF ノードに障害が発生しても接続には影響しません。

---

図 3-4 は、フェイルオーバーおよびスケーラブルの各リソースグループの例と、スケーラブルサービスでそれらの間に存在する依存関係を示しています。この例は、3つのリソースグループを示しています。フェイルオーバーリソースグループには、可用性の高い DNS のアプリケーションリソースと、可用性の高い DNS および可用性の高い Apache Web Server の両方によって使用されるネットワークリソースが含まれます。スケーラブルリソースグループには、Apache Web Server のアプリケーションインスタンスだけが含まれます。リソースグループの依存関係は、スケーラブルリソースグループとフェイルオーバーリソースグループの間に存在し(実線)、Apache アプリケーションリソースはすべて、共有アドレスであるネットワークリソース schost-2 に依存する(破線)ことに注意してください。

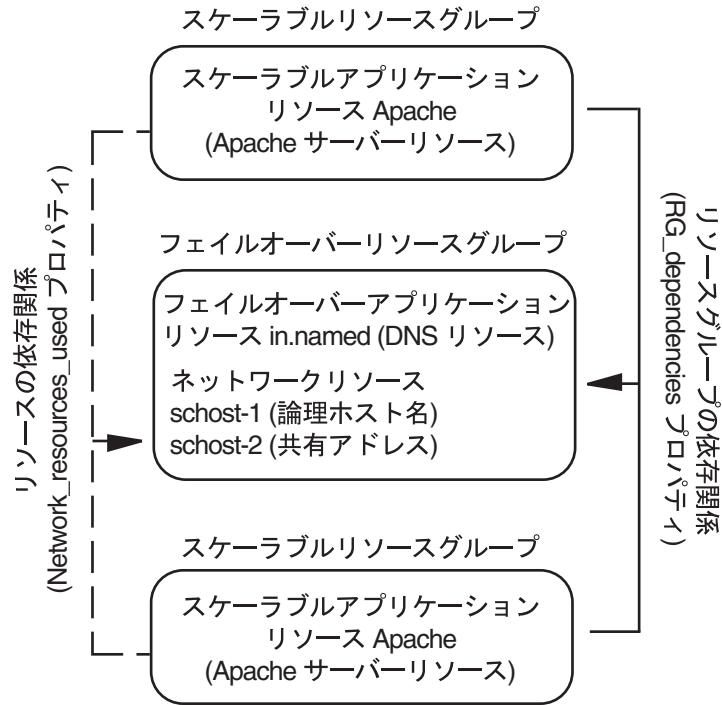


図 3-4 フェイルオーバーリソースグループとスケーラブルリソースグループの例

### スケーラブルサービスの構造

クラスタネットワークの主な目的は、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ともに、そのサービスの遠隔クライアントの観点から見て1つのサービスとして動作し、サービスの機能を実現します。たとえば、いくつかのノードで実行されるいくつかの httpd デーモンからなるスケーラブル Web サービスがあります。どの httpd デーモンもクライアント要求に対応できます。要求に対応するデーモンは、負荷均衡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ではなく、サービスによる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。

スケーラブルサービスは、次のものからなります。

- スケーラブルサービスに対するネットワークインフラストラクチャのサポート
- 負荷均衡
- ネットワーキングおよびデータサービスに対する HA サポート (リソースグループマネージャーを使用)

次の図は、スケーラブルサービスの構造を示しています。

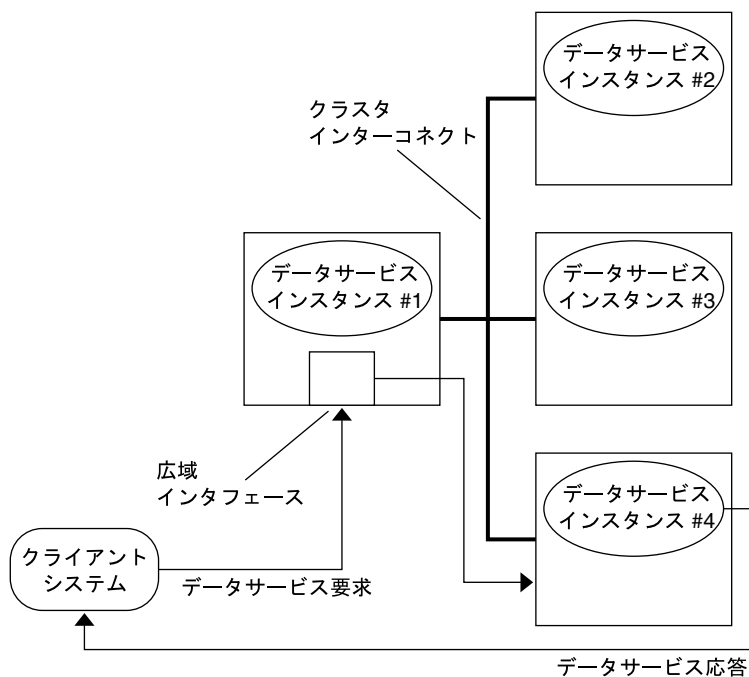


図 3-5 スケーラブルサービスの構造

広域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。GIF に入るパケットは、構成可能な負荷均衡ポリシーに基づいて、他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、そのいずれかのインスタンスがクライアント要求に応答できるサービスをいいます。*sticky* サービスとは、クライアントが同じインスタンスに要求を送るサービスをいいます。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、3 ノードクラスタでは、各ノードに 1 のウェイトがあるものと想定します。各ノードは、そのサービスに代わって、クライアントからの要求の 3 分の 1 のサービスを提供します。ウェイトは、`scrgadm(1M)` コマンドインタフェースを使用し、管理者がいつでも変更できます。

*sticky* サービスには、*ordinary sticky* と *wildcard sticky* の 2 種類があります。*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。単一ポートを待機しているそのサーバーインスタンスという点で、そのクライアントは *sticky* であると呼ばれます。クライアントは、インスタンスが起動してアクセス可能であり、負荷分散ポリシーがサーバー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。

たとえば、クライアント上の Web ブラウザは、3 つの異なる TCP 接続を使用して、ポート 80 にある共有 IP アドレスに接続しますが、これらの接続はサービスでキャッシュされたセッション情報を交換します。

*sticky* ポリシーを一般化すると、同じインスタンスの背後でセッション情報を交換する複数のスケーラブルサービスにまで及びます。これらのサービスが同じインスタンスの背後でセッション情報を交換する場合、同じノードで異なるポートと通信

する複数のサーバーインスタンスという点で、そのクライアントは *sticky* と呼ばれます。

たとえば、電子商取引サイトの顧客は、ポート 80 の HTTP を使用して買い物をしますが、購入した製品の支払いをクレジットカードで行うためには、ポート 443 で SSL に切り替えて機密データを送ります。

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やはり同じノードに送られるものと想定します。クライアントは、同じ IP アドレスという点で、ポートに対して *sticky wildcard* です。

このポリシーの例としては、受動モード FTP があります。クライアントは、ポート 21 の FTP サーバーに接続して、動的ポート範囲のリスナーポートサーバーに接続するよう、そのサーバーによって通知されます。この IP アドレスに対する要求はすべて、サーバーが制御情報によってクライアントに通知した、同じノードに転送されます。

これらの各 *sticky* ポリシーでは、ウェイト設定した (*weighted*) 負荷均衡ポリシーがデフォルトで有効であるため、クライアントの最初の要求は、負荷均衡によって指定されたインスタンスにリダイレクトされます。インスタンスが実行されているノードとクライアントが関係を確立すると、そのノードがアクセス可能で、負荷分散ポリシーが変更されない限り、今後の要求はそのインスタンスに送られます。

次に、各負荷均衡ポリシーの詳細について説明します。

- *weighted* - 負荷は指定されたウェイト値に従って各種のノードに分配されます。このポリシーは `Load_balancing_weights` プロパティに設定された `LB_WEIGHTED` の値を使用して設定されます。ウェイトがノードについて明示的に設定されていない場合は、デフォルトで 1 が設定されます。

このポリシーは、ラウンドロビンではないことに注意してください。ラウンドロビンポリシーでは、クライアントからの各要求が、最初の要求はノード 1、2 番目の要求はノード 2 といったように常に異なるノードに送られます。ウェイトを設定したポリシーは、一定の割合のクライアントトラフィックが、特定ノードに送られるようにするものです。このポリシーは、個々の要求には対応しません。

- *sticky* - このポリシーでは、ポートの集合が、アプリケーションリソースの構成時に認識されます。このポリシーは、`Load_balancing_policy` リソースプロパティの `LB_STICKY` の値を使用して設定されます。
- *sticky-wild* - このポリシーは、通常の *sticky* ポリシーの上位セットです。IP アドレスによって識別されるスケーラブルサービスでは、ポートはサーバーによって割り当てられます (したがって前もって認識されません)。ポートは変更されるこ



とがあります。このポリシーは、Load\_balancing\_policy リソースプロパティの LB\_STICKY\_WILD の値を使用して設定されます。

## フェイルバック設定

リソースグループは、ノードからノードへ処理を継続します。リソースグループが処理を別のノードに継続する場合、以前実行していたノードがクラスタに戻ってから、元のノードにフェイルバックするように指定できます。このオプションは、Failback リソースグループプロパティによって設定できます。

特定のインスタンスでは、たとえば、リソースグループをホストする元のノードに障害が生じて繰り返し再起動する場合、フェイルバックを設定すると、リソースグループの可用性が低下することがあります。

## データサービス障害モニター

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探索してその状態を判断する障害モニターがあります。障害モニターは、アプリケーションデーモンが実行されていて、クライアントにサービスが提供されていることを確認します。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。

## 新しいデータサービスの開発

Sun は、各種のアプリケーションを可用性の高いデータサービスとしてクラスタ内で動作させるためのソフトウェアを提供しています。可用性の高いデータサービスとして実行したいアプリケーションが、現在 Sun から提供されていない場合は、API または DSDL API を使用して、各自のアプリケーションを可用性の高いデータサービスとして実行するように構成できます。データサービスには、フェイルオーバーとスケラブルの2つの種類があります。アプリケーションがどちらのデータサービスを使用できるかを判断するための基準があります。個々の基準は、アプリケーションで使用できる API について説明している Sun Cluster のマニュアルに記載されています。

次に、各自のサービスがスケラブルデータサービスの構造を利用できるかどうかを知るために役立つガイドラインをいくつか示します。スケラブルサービスの一般的な情報については、60ページの「スケラブルデータサービス」を参照してください。

次のガイドラインを満たす新しいサービスは、スケーラブルサービスを利用できます。既存のサービスがこれらのガイドラインに従っていない場合は、そのサービスがガイドラインに準拠するように、一部を書き直さなければならない場合があります。

スケーラブルデータサービスには、以下の特性があります。まず、このようなサービスは、1つまたは複数のサーバーインスタンスからなります。各インスタンスは、クラスタの異なるノードで実行されます。同じサービスの複数のインスタンスを、同じノードで実行することはできません。

次に、サービスが外部論理データ格納を使用する場合は、この格納に対する複数のサーバーインスタンスからの同時アクセスの同期をとって、更新が失われたり、変更中のデータを読み取ったりすることを避ける必要があります。この格納をメモリー内部の状態と区別するために「外部」と呼び、格納がそれ自体複製されている場合でも単一の実体として見えるため、「論理的」と呼んでいることに注意してください。また、この論理データ格納には、サーバーインスタンスが格納を更新するたびに、その更新がすぐに他のインスタンスで見られるという特性があります。

Sun Cluster は、このような外部記憶領域をそのクラスタファイルシステムと広域 raw パーティションを介して提供します。例として、サービスが外部ログファイルに新しいデータを書き込む場合や既存のデータを修正する場合を想定してください。このサービスの複数インスタンスが実行されている場合は、それぞれがこの外部ログへのアクセスを持ち、同時にこのログにアクセスできます。各インスタンスは、このログに対するアクセスの同期をとる必要があります。そうしないと、インスタンスは相互に妨害しあうこととなります。サービスは、`fcntl(2)` および `lockf(3C)` によって通常の Solaris ファイルロックを使用して、必要な同期をとることができます。

このような格納のもう1つの例としては、可用性の高い Oracle または Oracle Parallel Server などのバックエンドデータベースが挙げられます。このようなバックエンドデータベースサーバーは、データベース照会または更新トランザクションを使用するのに内部組み込みの同期を使用するため、複数のサーバーインスタンスが独自の同期を実装する必要がありません。

現在の実現状態ではスケーラブルサービスではないサービスの例として、Sun の IMAP サーバーがあります。このサービスは格納を更新しますが、その格納はプライベートであり、複数の IMAP インスタンスがこの格納に書き込むと、更新の同期がとられないために相互に上書きし合うこととなります。IMAP サーバーは、同時アクセスの同期をとるために書き直す必要があります。

最後に、インスタンスは、他のインスタンスのデータから切り離されたプライベートデータを持つ場合があることに注意してください。このようなケースでは、デー

タはプライベートであり、そのインスタンスだけがデータを処理するため、サービスは同時アクセスの同期をとる必要はありません。この場合、このプライベートデータが広域にアクセス可能になる可能性があるため、このデータをクラスタファイルシステムのもとで保存しないように注意する必要があります。

## データサービス API と DSDL API

Sun Cluster には、アプリケーションの可用性を高めるための次のものがあります。

- Sun Cluster の一部として提供されるデータサービス
- データサービス API
- DSDL API (データサービス開発ライブラリ API)
- 汎用データサービス

『Sun Cluster 3.0 データサービスのインストールと構成』は、Sun Cluster によって提供されるデータサービスをインストールして構成する方法を説明しています。

『Sun Cluster 3.0 データサービス開発ガイド』は、Sun Cluster フレームワークのもとで他のアプリケーションの可用性を高くする方法を説明しています。

Sun Cluster API と DSDL API を使用すると、アプリケーションプログラマは、障害モニターおよびデータサービスインスタンスを起動して停止するスクリプトを開発できます。これらのツールを使用すると、アプリケーションをフェイルオーバーまたはスケラブルデータサービスとして設計できます。また、Sun Cluster には、アプリケーションに必要な起動および停止方式を簡単に生成して、可用性の高いデータサービスとして実行できるようにするための汎用データサービスもあります。

## リソースとリソースタイプ

データサービスは、数種類のリソースを使用します。つまり、Apache Web Server や iPlanet Web Server などのアプリケーションと、それらのアプリケーション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) です。アプリケーションとネットワークリソースは、RGM によって管理される基本ユニットを形成します。

リソースとは、クラスタ全体で定義されたリソースタイプをインスタンス化したものです。いくつかのリソースタイプが定義されています。

データサービスはリソースタイプです。たとえば、Sun Cluster HA for Oracle はリソースタイプ SUNW.oracle であり、Sun Cluster HA for Apache はリソースタイプ SUNW.apache です。

ネットワークリソースは、SUNW.LogicalHostname または SUNW.SharedAddress リソースタイプです。これら 2 つのリソースタイプは、Sun Cluster 製品により事前に登録されています。

SUNW.HAStorage リソースタイプは、リソースとそのリソースが依存するディスクデバイスグループの起動を同期させるのに使用します。これによって、データサービスが起動する前に、クラスタファイルシステムのマウントポイントへのパス、広域デバイス、デバイスグループ名を使用できます。

RGM 管理のリソースは、リソースグループと呼ばれるグループに入れられるため、1 つのユニットとして管理できます。リソースグループは、フェイルオーバーまたはスイッチオーバーがリソースグループで開始されたときに、ユニットで移行されます。

---

注 - アプリケーションリソースを含むリソースグループをオンラインにすると、アプリケーションが起動します。データサービス起動方式は、アプリケーションが起動して実行されるまで待ってから、正常に終了します。アプリケーションが起動して実行しているかどうかの判断は、データサービス障害モニターが、データサービスがクライアントにサービスを提供しているかどうかを判断するのと同じ方法で行われます。このプロセスの詳細については、『Sun Cluster 3.0 データサービスのインストールと構成』を参照してください。

---

## リソースとリソースグループプロパティ

Sun Cluster データサービスのリソースおよびリソースグループのプロパティ値は構成できます。一連の標準的なプロパティはすべてのデータサービスに共通であり、拡張プロパティは各データサービスに特定のものです。標準プロパティおよび拡張プロパティのいくつかは、デフォルト設定によって構成されているため、これらを修正する必要はありません。それ以外のプロパティは、リソースを作成して構成するプロセスの一部として設定する必要があります。各データサービスのマニュアルでは、リソースタイプで使用されるプロパティの種類とその構成方法を指定しています。

標準プロパティは、通常特定のデータサービスに依存しないリソースおよびリソースグループプロパティを構成するために使用されます。これらの標準プロパティについては、『Sun Cluster 3.0 データサービスのインストールと構成』の付録を参照してください。

拡張プロパティは、アプリケーションバイナリの場所や構成ファイルなどの情報を提供するものです。拡張プロパティは、データサービスの構成に従って修正する必

要があります。拡張プロパティについては、『Sun Cluster 3.0 データサービスのインストールと構成』のデータサービスに関する各章を参照してください。

## パブリックネットワーク管理 (PNM) とネットワークアダプタフェイルオーバー (NAFO)

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。各クラスタノードは、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。

Sun Cluster パブリックネットワーク管理 (PNM) 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して、障害が検出されたときにアダプタからアダプタへ IP アドレスの処理を継続するための基本機構を提供します。各クラスタノードには独自の PNM 構成があり、これは他のクラスタノードと異なることもあります。

パブリックネットワークアダプタは、ネットワークアダプタフェイルオーバーグループ (NAFO グループ) の中に構成されています。各 NAFO 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。特定の NAFO グループで一度にアクティブにできるのは 1 つのアダプタだけですが、同じグループ内の多くのアダプタは、アクティブアダプタの PNM デーモンによって障害が検出された場合のアダプタフェイルオーバーに使用されるバックアップアダプタとして機能します。フェイルオーバーでは、アクティブアダプタに関連する IP アドレスがバックアップアダプタに移動するため、ノードのパブリックネットワーク接続が維持されます。フェイルオーバーは、アダプタインタフェースレベルで発生するため、フェイルオーバー中に一時的にわずかな遅延がみられる以外、TCP などのこれよりも高いレベルの接続は影響されません。

---

**注** - TCP の構成回復特性が原因で、正常なフェイルオーバーの後、セグメントのいくつかがフェイルオーバー中に失われて、TCP の混雑制御機構をアクティブ化するために、TCP エンドポイントではさらに遅延が生じる可能性があります。

---

NAFO グループには、論理ホスト名と共有アドレスリソースの構築ブロックがあります。scrgadm(1M) コマンドは、必要に応じて NAF グループを自動的に作成します。論理ホスト名と共有アドレスリソースとは別に NAFO グループを作成して、クラスタノードのパブリックネットワーク接続を監視する必要もあります。ノード上の同じ NAFO グループが、任意の数の論理ホスト名、または共有アドレスリソースのホストとなることができます。論理ホスト名と共有アドレスリソースの詳細

については、『Sun Cluster 3.0 データサービスのインストールと構成』を参照してください。

---

注 - NAFO 機構の設計は、アダプタの障害を検出してマスクすることを目的としています。この設計は、`ifconfig(1M)` を使用して論理(または共有) IP アドレスのどれかを削除した状態から回復することを目的としていません。Sun Cluster の設計では、RGM によって管理されるリソースとして論理および共有 IP アドレスとみなします。管理者が IP アドレスを追加または削除する正しい方法は、`scrgadm(1M)` を使用してリソースを含むリソースグループを修正するというものです。

---

## PNM 障害検出とフェイルオーバープロセス

PNM は、正常なアダプタのパケットカウンタが、アダプタを介した通常のネットワークトラフィックによって変化するものと想定して、アクティブアダプタのパケットカウンタを定期的にチェックします。パケットカウンタがしばらくの間変化しない場合、PNM は ping シーケンスに入って、トラフィックを強制的にアクティブアダプタに送ります。PNM は、各シーケンスの最後でパケットカウンタに変化がないかを検査し、ping シーケンスが何度か繰り返された後でもパケットカウンタに変化がない場合は、アダプタの障害を宣言します。これらのイベントは、バックアップアダプタがあれば、それへのフェイルオーバーを引き起こします。

入力および出力パケットカウンタはいずれも PNM によって監視され、どちらかまたは両方にしばらくの間変化がない場合は、ping シーケンスが開始されます。

ping シーケンスは、ALL\_ROUTER マルチキャストアドレス (224.0.0.2)、ALL\_HOST マルチキャストアドレス (224.0.0.1)、およびローカルサブネットブロードキャストアドレスの ping からなります。

ping は、コストの低いもの順という方法で構成されているため、コストのかかる ping は、コストの低い ping が成功した場合は実行されません。また、ping はアダプタでのトラフィックを生成するための方法としてのみ使用されます。その終了状態は、アダプタが機能しているか障害が発生しているかの判断には関係ありません。

`inactive_time`、`ping_timeout`、`repeat_test`、`slow_network` の 4 つの調整可能なパラメータがこのアルゴリズムにあります。これらのパラメータによって、フォルト検出の速度と正確さを調整できます。パラメータの詳細とその変更方法については、『Sun Cluster 3.0 のシステム管理』のパブリックネットワークパラメータの変更手順を参照してください。

NAFO グループのアクティブアダプタで障害が検出された後で、バックアップアダプタが使用できない場合、そのグループは停止と宣言されますが、そのバックアップアダプタのすべてのテストは続行します。また、バックアップアダプタが使用可能な場合は、そのバックアップアダプタに対してフェイルオーバーが発生します。論理アドレスとその関連フラグはバックアップアダプタに転送されて、障害の発生したアクティブアダプタは停止して切り離された状態に (unplumbed) になります。

IP アドレスのフェイルオーバーが正常に終了すると、自動的に ARP が送信されます。したがって、遠隔クライアントへの接続は維持されます。





## 頻繁に寄せられる質問 (FAQ)

---

この章では、Sun Cluster に関して最も頻繁に寄せられる質問に対する回答を示します。回答は、トピックごとに構成されています。

---

### 高可用性に関する FAQ

- 可用性の高いシステムとは何ですか。

Sun Cluster では、高可用性 (HA) を、通常サーバーシステムを使用不能にするような障害が発生した場合でも、クラスタがアプリケーションを起動して実行できる能力として定義しています。

- クラスタが高可用性を提供するプロセスは何ですか。

クラスタフレームワークは、フェイルオーバーとして知られるプロセスによって可用性の高い環境を提供します。フェイルオーバーとは、障害の発生したノードからクラスタ内の別の動作可能ノードにアプリケーションを移行するために、クラスタによって実行される一連のステップです。

- HA サービスとスケラブルサービスの違いは何ですか。

HA サービスとは、アプリケーションが一度に1つのクラスタ内の主ノードだけで実行されることを示します。他のノードは他のアプリケーションを実行できませんが、各アプリケーションは単一のノードでのみ実行されます。主ノードに障害が発生すると、障害が発生したノードで実行されていたアプリケーションは、別のノードに処理を引き継いで実行を続けます。

スケーラブルサービスは、アプリケーションを複数のノードに広げて、単一の論理サービスを作成します。スケーラブルサービスは、実行されるクラスタ全体のノードとプロセッサの数を強化します。1つのノードがすべてのアプリケーション要求を受け取り、アプリケーションサーバーが実行されている複数のノードにその要求を振り分けます。このノード(広域インタフェースノードまたは GIF と呼ばれる)に障害が発生すると、広域インタフェースは存続するノードに処理を引き継ぎます。アプリケーションが実行されているノードに障害が発生すると、アプリケーションは別のノードで実行を続けますが、障害が発生したノードがクラスタに戻るまで多少のパフォーマンス低下が生じます。

---

## ファイルシステムに関する FAQ

- 1つまたは複数のクラスタノードを、他のクラスタノードをクライアントとして使用して、可用性の高い **NFS** サーバーとして実行できますか。

実行できません。ローカルロックに関する問題が `lockd` の強制終了と再起動(NFS フェイルオーバー中に発生する)の機能を妨害します。終了と再起動の間で、阻止されたローカルプロセスにロックが付与される可能性があるため、ロックを所有するクライアントシステムがフェイルオーバー後にロックを再利用できなくなります。

- リソースグループマネージャー (**RGM**) によって制御されていないアプリケーションでクラスタファイルシステムを使用できますか。

使用できます。ただし、**RGM** の制御下にないと、アプリケーションは、実行されているノードに障害があった場合に存続できません。

- すべてのクラスタファイルシステムのマウントポイントが、`/global/device-group` ディレクトリになければなりませんか。

いいえ。ただし、クラスタファイルシステムを `/global/device-group` などの同一のマウントポイントのもとに置くと、これらのファイルシステムの構成と管理は簡単になります。

- クラスタファイルシステムを使用した場合と **NFS** ファイルシステムをエクスポートした場合の違いは何ですか。

次に示すいくつかの違いがあります。

1. クラスタファイルシステムは広域デバイスをサポートします。**NFS** は、デバイスへの遠隔アクセスをサポートしません。

2. クラスタファイルシステムには広域名前空間があります。したがって、必要なのは1つのマウントコマンドだけです。これに対し、NFSでは、ファイルシステムを各ノードにマウントする必要があります。
3. クラスタファイルシステムは、NFSよりも多くの場合でファイルをキャッシュします。たとえば、ファイルが、読み取り、書き込み、ファイルロック、非同期入出力のために複数のノードからアクセスされている場合です。
4. クラスタファイルシステムは、1つのサーバーに障害が生じた場合のシームレスなフェイルオーバーをサポートします。NFSは複数のサーバーをサポートしますが、フェイルオーバーは読み取り専用ファイルシステムにのみ可能です。
5. クラスタファイルシステムは、リモート DMA とゼロコピー機能を提供する、将来の高速クラスタインターコネクトを利用するために組み込まれています。
6. クラスタファイルシステムのファイルの属性を (`chmod(1M)` などを使用して) 変更すると、変更内容はすべてのノードでただちに反映されます。エクスポートされた NFS ファイルシステムでは、この処理に時間がかかる場合があります。

---

## ボリューム管理に関する FAQ

- すべてのディスクデバイスをミラー化する必要がありますか。

ディスクデバイスの可用性を高くするには、それをミラー化するか、RAID-5 ハードウェアを使用する必要があります。すべてのデータサービスは、可用性の高いディスクデバイスか、可用性の高いディスクデバイスにマウントされたクラスタファイルシステムのどちらかを使用する必要があります。このような構成では、単一のディスク障害に耐えられます。

---

## データサービスに関する FAQ

- 利用可能な **Sun Cluster** データサービスは何ですか。

サポートされているデータサービスについては、『*Sun Cluster 3.0* ご使用にあたって』を参照してください。

- **Sun Cluster** データサービスによってサポートされているアプリケーションのバージョンは何ですか。

サポートされているアプリケーションのバージョンについては、『*Sun Cluster 3.0 Error Messages Manual*』を参照してください。

- 独自のデータサービスを作成できますか。

できます。詳細については、『*Sun Cluster 3.0 データサービス開発ガイド*』と DSDL API と共に提供される文書「*Data Service Enabling Technologies*」を参照してください

- ネットワークリソースを作成する場合に、IP アドレスで指定するのですか。またはホスト名で指定するのですか。

ネットワークリソースを指定する場合には、IP アドレスではなく、UNIX のホスト名を使用することを推奨します。

- ネットワークリソースを作成する場合に、論理ホスト名 (**LogicalHostname** リソース) または共有アドレス (**SharedAddress** リソース) を使用した場合の違いは何ですか。

Failover モードリソースグループの LogicalHostname リソースを使用する場合は、SharedAddress リソースと LogicalHostname リソースは同様に使用できます。SharedAddress リソースを使用すると、クラスタネットワークングソフトウェアが LogicalHostname ではなく、SharedAddress に合わせて構成されているために、多少のオーバーヘッドが生じます。

SharedAddress を使用する利点は、スケーラブルおよびフェイルオーバーの両方のデータサービスを構成して、クライアントが同じホスト名を使用して両方のサービスにアクセスできるようにする場合に得られます。この場

合、SharedAddress リソースはフェイルオーバーアプリケーションリソースとともに 1 つのリソースグループに含まれますが、スケーラブルサービスリソースは個別のリソースグループに含まれ、SharedAddress を使用するよう構成されます。スケーラブルおよびフェイルオーバーサービスはどちら

も、SharedAddress リソースに構成された同じホスト名とアドレスのセットを使用します。

---

## パブリックネットワークに関する FAQ

- **Sun Cluster** がサポートするパブリックネットワークアダプタは何ですか。

現在、Sun Cluster は、Ethernet (10/100BASE-T および 1000BASE-SX Gb) パブリックネットワークアダプタをサポートしています。今後新しいインタフェース

がサポートされる可能性があるため、最新情報については、ご購入先に確認してください。

- フェイルオーバーでの **MAC** アドレスの役割は何ですか。

フェイルオーバーが発生すると、新しいアドレス解決プロトコル (ARP) パケットが生成されて伝送されます。これらの ARP パケットには、新しい MAC アドレス (ノードの処理が実行される新しい物理アダプタのアドレス) と古い IP アドレスが含まれます。ネットワーク上の別のマシンがこれらのパケットの1つを受け取ると、その ARP キャッシュから古い MAC-IP マッピングがフラッシュされて、新しいものが使用されます。

- **Sun Cluster** では、ホストアダプタ用の **OpenBoot PROM** に **local-mac-address?=true** を設定できますか。

設定できません。この変数はサポートされていません。

---

## クラスタメンバーに関する FAQ

- すべてのクラスタメンバーが同じ **root** パスワードを持つ必要がありますか。

各クラスタメンバーに同じ **root** パスワードを設定する必要はありません。ただし、同じ **root** パスワードをすべてのノードに使用すると、クラスタの管理を簡略化できます。

- ノードが起動される順序は重要ですか。

ほとんどの場合、重要ではありません。ただし、起動順序は、**amnesia** (詳細は、52ページの「定足数と定足数デバイス」を参照) を防止するために重要です。たとえば、ノード2が定足数デバイスの所有者であり、ノード1が停止してノード2を停止させた場合は、ノード2を起動してからノード1を起動する必要があります。これにより、古いクラスタ構成情報を持つノードを誤って起動するのを防ぐことができます。

- クラスタノードのローカルディスクをミラー化する必要がありますか。

必要ありません。このミラー化は必要条件ではありませんが、クラスタノードのディスクをミラー化すると、ノードを停止させる非ミラー化ディスクの障害を防止できます。ただし、クラスタノードのローカルディスクをミラー化すると、システム管理の負荷が増えます。

- クラスタメンバーのバックアップの注意点は何か。

クラスタには、いくつかのバックアップ方式を使用できます。1つの方法としては、ノードをテープドライブまたはライブラリが接続されたバックアップノードとして設定します。さらに、クラスタファイルシステムを使用してデータをバックアップします。このノードは共有ディスクには接続しないでください。

バックアップと復元の手順については、『Sun Cluster 3.0 のシステム管理』を参照してください。

---

## クラスタ記憶装置に関する FAQ

- 多重ホスト記憶装置の可用性を高めるものは何ですか。

多重ホスト記憶装置は、単一のディスクが失われてもミラー化(またはハードウェアベースの RAID-5 コントローラ)のために存続できるので、可用性が高くなります。多重ホスト記憶装置には複数のホスト接続があるため、接続先の単一ノードが失われても耐えることができます。

- サポートされている多重ホスト記憶装置構成は何ですか。

現在、2 ノードを超える接続はサポートされていません。単一の格納装置内にあるすべての多重ホストディスクが、同じ2つのノードに接続しなければなりません。詳細については、32ページの「Sun Cluster トポロジ」を参照してください。

- SCSI-3 PGR 用に構成されたディスクを広域デバイスとして使用できますか。

現在、SCSI-3 PGR は Sun Cluster ではサポートされていません。SCSI-2 方式だけが広域ディスクデバイスとしてサポートされています。SCSI-3 ディスクはサポートされていないため、`scdidadm(1M)` の `-R` オプションを使用して、クラスタで広域デバイスとして使用するすべての SCSI-3 ディスクに対して、正しい SCSI 方式を設定する必要があります。

---

## クラスタインターコネクトに関する FAQ

- Sun Cluster がサポートするクラスタインターコネクトは何ですか。

現在、Sun Cluster は、Ethernet (100BASE-T Fast Ethernet および 1000BASE-SX Gb) クラスタインターコネクトをサポートしています。Scalable Coherent Interface (SCI) のサポートも計画されています。

---

## クライアントシステムに関する FAQ

- クラスタでの使用における特殊なクライアントの要求や制約について考慮する必要がありますか。

クライアントシステムは、他のサーバーに接続する場合と同様にクラスタに接続します。データサービスアプリケーションによっては、クライアント側ソフトウェアをインストールするか、別の構成変更を行なって、クライアントがデータサービスアプリケーションに接続できるようにしなければならないこともあります。クライアント側の構成条件の詳細については、『Sun Cluster 3.0 データサービスのインストールと構成』を参照してください。

---

## 管理コンソールに関する FAQ

- Sun Cluster には管理コンソールが必要ですか。

必要です。

- 管理コンソールをクラスタ専用にする必要がありますか、または別の作業に使用することができますか。

Sun Cluster は、専用の管理コンソールを必要としません。ただし、専用のコンソールを使用すると、次の利点が得られます。

- コンソールと管理ツールを同じマシンにグループ化することによって、集中的なクラスタ管理を行うことができます。
- ハードウェアの保守担当者による迅速な問題解決が可能になります。
- 管理コンソールをクラスタの近く、たとえば同じ部屋に配置する必要がありますか。  
ハードウェアの保守担当者に確認してください。保守作業上、コンソールをクラスタの近くに配置する必要がある場合があります。コンソールを同じ部屋に配置する必要性は、技術的にはありません。
- 距離の条件すべてが合致するかぎり、1台の管理コンソールが複数のクラスタにサービスを提供できますか。

提供できます。複数のクラスタを1台の管理コンソールから制御できます。また、1台の端末集配信装置(コンセントレータ)をクラスタ間で共有することもできます。

---

## 端末集配信装置とシステムサービスプロセッサに関する FAQ

- **Sun Cluster** は端末集配信装置を必要としますか。

Sun Cluster 3.0 を実行するために端末集配信装置は必要はありません。障害による影響防止に端末集配信装置を必要とした Sun Cluster 2.2 とは異なり、Sun Cluster 3.0 は端末集配信装置に依存しません。

- ほとんどの **Sun Cluster** サーバーは端末集配信装置を使用していますが、**E10000** は使用していないのはなぜですか。

端末集配信装置は、ほとんどのサーバーで効率的なシリアル - Ethernet コンバータです。そのコンソールポートはシリアルポートです。Sun Enterprise E10000 サーバーには、シリアルコンソールがありません。システムサービスプロセッサ (SSP) は Ethernet または jtag ポートを介したコンソールです。Sun Enterprise E10000 サーバーの場合は、コンソールに対して常に SSP を使用します。

- 端末集配信装置を使用した場合の利点は何ですか。

端末集配信装置を使用すると、ノードが OpenBoot PROM (OBP) にある場合でも、ネットワーク上の任意の場所にある遠隔ワークステーションから各ノードに対して、コンソールレベルのアクセスを行えます。

- **Sun** がサポートしていない端末集配信装置を使用する場合に注意する点は何ですか。

Sun がサポートする端末集配信装置と他のコンソールデバイスの主な違いは、Sun の端末集配信装置には、端末集配信装置がコンソールに対して起動時にブレイクを送信するのを防ぐ特殊なファームウェアがあるという点です。ブレイク、またはコンソールに対してブレイクと解釈されることがある信号を送信するコンソールデバイスである場合、ノードが停止されることに注意してください。

- **Sun** がサポートする端末集配信装置を再起動しないで、そこにあるロックされたポートを開放できますか。

開放できます。リセットする必要があるポート番号を書きとめて、次のコードを実行してください。



```
telnet tc
Enter Annex port name or number: cli
annex: su -
annex# admin
admin : reset port_number
admin : quit
annex# hangup
#
```

Sun がサポートする端末集配信装置の構成と管理の詳細については、『*Sun Cluster 3.0* のシステム管理』を参照してください。

- 端末集配信装置自体に障害が発生した場合はどうしたらいいですか。別の装置を用意しておく必要がありますか。

ありません。端末集配信装置に障害が発生しても、クラスタの可用性はまったく失われません。端末集配信装置が再び機能するまで、ノードコンソールに接続できなくなります。

- 端末集配信装置を使用する場合に、セキュリティはどのように制御しますか。

通常、端末集配信装置は、他のクライアントアクセスに使用されるネットワークではなく、システム管理者が使用する小規模なネットワークに接続されています。この特定のネットワークに対するアクセスを制限することでセキュリティを制御できます。



# 用語集

---

この以下の用語は、Sun Cluster 3.0 のマニュアルで使用されています。

## A

**amnesia** クラスタの停止後そのクラスタが無効なクラスタ構成データ (CCR) を使用して再起動する状態。たとえば、ノード 1 だけが動作可能な 2 ノードクラスタで、クラスタ構成の変更がノード 1 で発生した場合、ノード 2 の CCR は無効になります。クラスタ全体が停止してノード 2 から再起動した場合、ノード 2 の無効な CCR が全体原因で amnesia 状態が発生する。

## D

**DID ドライバ** Sun Cluster によって実装されるドライバで、クラスタ間で一貫したデバイス名前空間を提供する。「DID 名」も参照。

**DID 名** Sun Clusterの広域デバイスを識別するために使用される。Solaris 論理名と 1 対 1 または 1 対多の関係を持つクラスタ化識別子で、dXsY の形式をとる。X は整数、Y はスライス名を示す。「Solaris 論理名」も参照。

## H

**HA データサービス** 「データサービス」を参照。

## S

<b>Scalable Coherent Interface (SCI)</b>	クラスタインターコネクトとして使用される高速インターコネクトハードウェア。
<b>Solaris</b> 物理名	Solarisのデバイスドライバによってデバイスに指定される名前。これは、Solaris マシンを、/devices ツリーのパスとして示す。たとえば、典型的な SCSI ディスクには、次のような Solaris 物理名がある。 <code>/devices/sbus@1f,0/SUNW,fas@e,8800000/sd@6,0:c,raw</code> 「Solaris 論理名」も参照。
<b>Solaris</b> 論理名	通常、Solaris デバイスの管理に使用される名前。ディスクの場合、これらは通常、/dev/rdisk/c0t2d0s2 のようになる。これらの各 Solaris 論理デバイス名には、対応する Solaris 物理デバイス名がある。「DID 名」と「Solaris 物理名」も参照。
<b>Solstice DiskSuite</b>	Sun Cluster によって使用されるボリューム管理ソフトウェア。「ボリューム管理ソフトウェア」も参照。
<b>split brain</b>	クラスタが複数のパーティションに分割される状態。各パーティションは、相互の存在を認識しないで形成される。

## V

<b>VERITAS Volume Manager</b>	Sun Cluster によって使用されるボリューム管理ソフトウェア。「ボリューム管理ソフトウェア」も参照。
-------------------------------	--

## い

イベント	管理されるオブジェクトの状態、マスター、重大度の説明における変更。
インスタンス	「リソース起動」を参照。

## か

管理コンソール クラスタ管理ソフトウェアの実行に使用されるワークステーション。

## き

共有アドレスリソース クラスタ内のノードで実行されるすべてのスケラブルサービスによって結合されるネットワークアドレス。これによりノード上のサービスがスケラブルになる。クラスタは複数の共有アドレスを持つことができ、サービスは複数の共有アドレスに結合できる。

## く

クラスタ クラスタファイルシステムを共有し、フェイルオーバー、パラレルリソース、またはスケラブルリソースを実行するためにともに構成された、複数の相互接続されたノードやドメイン。

クラスタインターコネク ト ケーブル、クラスタトランスポート接続点、およびクラスタトランスポートアダプタから成るハードウェアのネットワークインフラストラクチャ。Sun Cluster およびデータサービスソフトウェアにより、クラスタ内通信で使用される。

クラスタ構成レポジトリ (CCR) Sun Cluster ソフトウェアによって、クラスタ構成情報を持続的に保存するために使用される可用性の高い複製データストア。

クラスタトランスポートアダプタ ノードに常駐して、ノードをクラスタインターコネク トに接続するネットワークアダプタ。「クラスタインターコネク ト」も参照。

クラスタトランスポートケーブル 終端に接続するネットワーク接続。クラスタトランスポートアダプタとクラスタトランスポート接続点の間の接続、または2つのクラスタトランスポートアダプタの間の接続。「クラスタインターコネク ト」も参照。

クラスタトランスポート接続点 クラスタインターコネク トの一部として使用されるハードウェアスイッチ。「クラスタインターコネク ト」も参照。

クラスタメンバー 現在のクラスタ実装のアクティブメンバー。このメンバーは、他のクラスタメンバーとリソースを共有して、他のクラスタメン

バー、およびクラスタのクライアントの両方にサービスを提供できる。「クラスタノード」も参照。

クラスタノード	クラスタメンバーとして構成されたノード。クラスタノードは、現在のメンバーである場合も、そうでない場合もある。「クラスタメンバー」も参照。
クラスタファイルシステム	既存のローカルファイルシステムに対し、クラスタ全体で可用性の高いアクセスを提供するクラスタサービス。
クラスタメンバーシップモニター (CMM)	クラスタメンバーシップの一貫性を維持するソフトウェア。このメンバーシップ情報は、残りのクラスタ化されたソフトウェアによって、可用性の高いサービスを配置する場所を決定するために使用される。CCM は、非クラスタメンバーがデータを破壊したり、破壊されたデータや矛盾したデータをクライアントに送信できないようにする。

ハ

広域インタフェース	共有アドレスの物理的なホストとなる広域ネットワークインタフェース。「共有アドレス」も参照。
広域インタフェースノード	広域インタフェースのホストとなるノード。
広域デバイス	ディスク、CD-ROM、テープなど、すべてのクラスタメンバーからアクセスできるデバイス。
広域デバイス名前空間	広域デバイスを示すクラスタ全体にわたる論理的な名前空間。 Solaris 環境のローカルデバイスは、/dev/dsk、/dev/rdsk、/dev/rmt ディレクトリに定義されている。広域デバイス名前空間は、広域デバイスを、/dev/global/dsk、/dev/global/rdsk、/dev/global/rmt の各ディレクトリに定義する。
広域リソース	Sun Cluster ソフトウェアのカーネルレベルで提供される可用性の高いリソース。広域リソースには、ディスク (HA デバイスグループ)、クラスタファイルシステム、広域ネットワークングを含めることができる。

コロケーション  
(collocation) 同じノードに存在するプロパティ。この概念は、パフォーマンスを向上させるために、クラスタ構成中に使用される。

## し

システムサービスプロセッサ (SSP) Sun Enterprise 10000 構成で、特にクラスタメンバーとの通信に使用されるクラスタ外部のデバイス。

自動フェイルバック 主ノードがエラーを起こし、その後にクラスタのメンバーとして再起動された場合、リソースグループまたはデバイスグループをその主ノードに返すプロセス。

終端 クラスタトランスポートアダプタまたはクラスタトランスポート接続点上の物理ポート。

主ノード リソースグループまたはデバイスグループが現在オンラインであるノード。つまり、主ノードとは、現在リソースに関連するサービスをホストしているか、または実装しているノードをいう。「二次ノード」も参照。

主ホスト名 主パブリックネットワーク上のノードの名前。常に、`/etc/nodename` に指定されたノード名になる。「二次ホスト名」も参照。

障害モニター データサービスの各部分を検出して、処置をとるために使用される障害デーモンとプログラム。「リソースモニター」も参照。

## す

スイッチオーバー クラスタ内のあるマスター (ノード) から別のマスター (または、リソースグループが複数の主ノードに構成されている場合は複数のマスター) へのリソースグループ、またはデバイスグループの正常な転送。スイッチオーバーは、管理者によって、`scswitch(1M)` コマンドを使用して開始される。

スイッチバック 「フェイルバック」を参照。

スケーラブルサービス	複数のノードで同時に実行されるように実装されたデータサービス。
スケーラブルリソース	複数のノードで実行されるリソース (各ノードのインスタンス) で、クラスタインターコネクトを使用して、サービスの遠隔クライアントに対して単一サービスの外観を提示する。

## せ

潜在主ノード	主ノードに障害が発生した場合に、フェイルオーバーリソースをマスターできるクラスタメンバー。「デフォルトマスター」も参照。
潜在マスター	「潜在主ノード」を参照。

## た

多重ホームホスト	複数のパブリックネットワーク上にあるホスト。
多重ホストディスク	物理的に複数のノードに接続されたディスク。
単一インタフェースリソース	クラスタ全体で最大で 1 つのリソースをアクティブにできるリソース。
端末集配信装置	Sun Enterprise 10000 以外の構成で、特にクラスタメンバーとの通信に使用されるクラスタ外部のデバイス。コンセントレータとも言う。

## ち

チェックポイント	ソフトウェアの状態の同期を取るよう指示するため、主ノードから二次ノードに対して送信される通知。「主ノード」と「二次ノード」も参照。
----------	---

## て

テイクオーバー	「フェイルオーバー」を参照。
---------	----------------



ディスクデバイスグループ	「デバイスグループ」を参照。
ディスクグループ	「デバイスグループ」を参照。
ディスクセット	「デバイスグループ」を参照。
定足数デバイス	定足数 ( <b>quorum</b> ) を確立してクラスタを実行するために使用される票を持つ、複数のノードによって共有されるディスク。クラスタは、定足数の票が利用可能な場合にのみ動作する。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て、新しいクラスタを構成するノードの集合を設定するときに使用される。
データサービス	リソースグループマネージャ ( <b>RGM</b> ) の制御下で、可用性の高いリソースとして実行されるように実装されたアプリケーション。
デバイス ID	Solaris を介して使用可能になるデバイスを識別する機構。 <code>devid_get(3DEVID)</code> のマニュアルページも参照。  Sun Cluster DID ドライバは、デバイス ID を使用して、異なるクラスタノードの Solaris 論理名間の相関関係を判断する。DID ドライバは、各デバイスでそのデバイス ID を検証する。そのデバイス ID が、クラスタ内の他の場所にある別のデバイスと一致する場合、両方のデバイスに同じ DID 名が指定される。そのデバイス ID がクラスタ内にない場合は、新しい DID が割り当てられる。 「Solaris 論理名」と「DID ドライバ」も参照。
デバイスグループ	クラスタ HA 構成内の異なるノードからマスター可能な、ディスクなどのデバイスリソースのユーザー定義グループ。このグループには、ディスクのデバイスリソース、Solstice DiskSuite ディスクセット、および VERITAS Volume Manager ディスクグループを含めることができる。
デフォルトマスター	フェイルオーバーリソースタイプがオンラインになるデフォルトクラスタメンバー。

## に

二次ノード	主ノードに障害が発生した場合にディスクデバイスグループおよびリソースグループをマスターするために使用できるクラスタメンバー。「主ノード」も参照。
二次ホスト名	二次パブリックネットワークのノードにアクセスするために使用される名前。「主ホスト名」も参照。

## ね

ネットワークアダプタフェイルオーバー (NAFO) グループ	同じノード、およびアダプタの障害時に相互にバックアップするように構成された同じサブネットにある 1 つまたは複数のネットワークアダプタのセット。
ネットワークアドレスリソース	「ネットワークリソース」を参照。
ネットワークリソース	1 つまたは複数の論理ホスト名または共有アドレスを含むリソース。「論理ホスト名リソース」と「共有アドレスリソース」も参照。

## の

ノード	Sun Cluster の一部にできる物理的なマシンまたは (Sun Enterprise E10000 サーバーの) ドメイン。「ホスト」とも呼ばれる。
-----	---

## は

ハートビート	利用可能なクラスタインターコネクトトランスポートパスすべてに送信される定期的なメッセージ。指定の間隔および再試行回数後にハートビートがない場合は、別のパスに対するトランスポート通信の内部フェイルオーバーが引き起こされる。クラスタメンバーへのすべてのパスに障害が発生した場合は、CMM によるクラスタ定足数 (quorum) の再評価が行われる。
--------	--

バックアップグループ	「ネットワークアダプタフェイルオーバーグループ」を参照。
パブリックネットワーク管理 (PNM)	障害モニターとフェイルオーバーを使用し、単一のネットワークアダプタやケーブルの障害が原因でノードの可用性が失われるのを防ぐソフトウェア。PNM フェイルオーバーは、ネットワークアダプタフェイルオーバーグループと呼ばれるネットワークアダプタの集合を使用し、クラスタノードとパブリックネットワークの間の冗長接続を実現する。障害モニターとフェイルオーバー機能は共に動作し、リソースの可用性を確保する。「ネットワークアダプタフェイルオーバーグループ」も参照。
パラレルリソースタイプ	クラスタ環境で実行されて、複数の (2 つ以上の) ノードによって同時にマスターできるように実装された、パラレルデータベースなどのリソースタイプ。
パラレルサービスインスタンス	個々のノードで実行されるパラレルリソースタイプのインスタンス。
汎用リソース	汎用リソースタイプの一部としてリソースグループマネージャの制御下におかれるアプリケーションデーモンとその子プロセス。
汎用リソースタイプ	データサービスのテンプレート。汎用リソースタイプは、単純なアプリケーションをフェイルオーバーデータサービス (あるノードで停止して、別のノードで起動する) にするために使用できる。このタイプは、Sun Cluster API によるプログラミングを必要としない。

## ひ

非クラスタモード	-x 起動オプションによってクラスタメンバーを起動した結果の状態。この状態では、ノードはクラスタメンバーではなくなるが、引き続きクラスタノードである。「クラスタメンバー」と「クラスタノード」も参照。
----------	---

## ふ

フェイルオーバー	障害発生後、現在の主ノードから新しい主ノードにリソースグループまたはデバイスグループを自動的に再配置すること。
----------	---

フェイルオーバーリソース	リソースの1つで、その各リソースは、一度に1つのノードによって正しくマスターできる。「単一インスタンスリソース」と「スケーラブルリソース」も参照。
フェイルバック	「自動フェイルバック」を参照。
フェイルファースト (failfast)	無効な操作によって損傷が判明する前に、障害が発生したノードのクラスタを正常に終了して移動すること。
負荷均衡	スケーラブルサービスにのみ適用される。アプリケーション負荷をクラスタ内のノード全体に分配して、クライアント要求が適宜サービスされるようにするプロセス。詳細については、60ページの「スケーラブルデータサービス」を参照。
負荷均衡ポリシー	スケーラブルサービスにのみ適用される。アプリケーションが要求した負荷をノード間で分配するための推奨される方法。詳細については、60ページの「スケーラブルデータサービス」を参照。
プライベートホスト名	クラスタインターコネクトを介してノードと通信するために使用されるホスト名のエイリアス。
分散ロックマネージャ (DLM)	共有ディスク Oracle Parallel Server (OPS) 環境で使用されるロックソフトウェア。DLMを使用すると、異なる複数のノードで実行される Oracle プロセスが、データベースアクセスの同期をとることができる。DLMは、高い可用性を目的として設計されている。プロセスまたはノードがクラッシュしても、残りのノードを停止して再起動する必要はない。DLMの高速再構成が実行されて、このような障害から回復する。

## ほ

ボリューム管理ソフトウェア	ディスクのストライプ化、連結、ミラー化、メタデバイスやボリュームの動的成長によってデータの信頼性を提供するソフトウェア製品。
---------------	--

## ま

マスター 「主ノード」を参照。

## め

メタデバイス状態  
データベースの複製  
(複製) ディスクに保存された、すべてのメタデバイスの構成と状態およびエラー条件を記録するデータベース。この情報は、Solstice DiskSuite ディスクセットの正しい操作に重要であり、複製される。

## り

リソース リソースタイプのインスタンス。同じタイプの多数のリソースが存在して、各リソースが独自の名前とプロパティ値を持つため、実際に使用するアプリケーションの多数のインスタンスをクラスタで実行できる。

リソース管理 API  
(RMAPI) クラスタ環境においてアプリケーションの可用性を高める、Sun Cluster のアプリケーションプログラミングインタフェース。

リソースグループ RGM によって 1 つの単位として管理されるリソースの集合。RGM によって管理される各リソースは、リソースグループに構成する必要がある。通常は、関連性があり相互に依存するリソースがグループ化される。

リソースグループ状態 ノードのリソースグループの状態。

リソースグループマネージャ (RGM) クラスタリソースを指定のクラスタノードで自動的に起動、停止することによって、これらのリソースの可用性を高め、スケラブルにするために使用されるソフトウェア機能。RGM は、構成済みのポリシーに従って、ハードウェアまたはソフトウェア障害、あるいは再起動時に動作する。

リソース状態 指定ノードのリソースグループマネージャリソースの状態。

リソースタイプ データサービス、LogicalHostname、SharedAddress クラスタオブジェクトに指定された一意の名前。データサービスリソースタイプは、フェイルオーバータイプかスケラブルタイプのどちら

かである。「データサービス」、「フェイルオーバーリソース」、「スケーラブルリソース」も参照。

リソースタイププロパティ	RGMによってリソースタイプの一部として保存され、指定タイプのリソースを記述して管理するために使用される重要な値のペア。
リソースステータス	障害モニターによってレポートされるリソースの状態。
リソースモニター	リソースが正常に実行されているかどうか、およびどのように実行されているかを判別するための定期的な障害検出をリソースに対して実行する、リソースタイプ実装のオプション部分。
リソース呼び出し	ノードで実行されるリソースタイプのインスタンス。ノードで起動されたリソースを表す抽象的な概念。

## ろ

ローカルディスク	特定のクラスタノードに物理的に接続された専用のディスク。
論理ネットワークインタフェース	インターネットアーキテクチャでは、ホストは1つまたは複数のIPアドレスを持つことができる。Sun Clusterは、追加の論理ネットワークインタフェースを構成して、いくつかの論理ネットワークインタフェースと単一の物理ネットワークインタフェース間のマッピングを確立する。各論理ネットワークインタフェースは単一のIPアドレスを持つ。このマッピングを使用すると、単一の物理ネットワークインタフェースが複数のIPアドレスに回答できる。また、このマッピングを使用すると、テイクオーバーやスイッチオーバーのときに、ハードウェアインタフェースを追加することなく、IPアドレスを1つのクラスタメンバーから別のメンバーに移動できる。
論理ホスト	アプリケーション、ディスクセット、またはアプリケーションデータが常駐するディスクグループと、クラスタへのアクセスに使用されるネットワークアドレスを含む、リソースのセットを表す Sun Cluster 2.0 の概念。この概念は、Sun Cluster 3.0 には存在しない。この概念を Sun Cluster 3.0 で実装する方法については、45ページの「ディスクデバイスグループ」と67ページの「リソースとリソースタイプ」を参照してください。

論理ホスト名リソース      ネットワークアドレスを表す論理ホスト名の集合を含むリソース。  
ス                              論理ホスト名リソースは、一度に1つのノードによってのみマ  
スターできる。「論理ホスト」も参照。