Sun Cluster 3.1 10/03 Konzepthandbuch

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

Teilenr.: 817–4258–10
November 2003, Version A

040106@7518

Inhalt

Vorwort

Das *Sun™ Cluster 3.1 Konzepthandbuch* enthält konzeptionelle und Verweisinformationen zum SunPlex™-System. Zum SunPlex-System gehören alle Hardware- und Softwarekomponenten, aus denen die Sun Cluster-Lösung besteht.

Dieses Dokument richtet sich an erfahrene Systemverwalter, die mit der Sun Cluster-Software vertraut sind. Verwenden Sie dieses Dokument weder als Planungs- noch als Presales-Unterlage. Vor der Lektüre dieses Dokuments sollten die Systemanforderungen feststehen, und Sie sollten im Besitz der geeigneten Geräte und Software sein.

Zum Verständnis der in diesem Buch beschriebenen Konzepte müssen Sie mit der Solaris™-Betriebsumgebung vertraut sein und Erfahrung mit der Datenträger-Manager-Software besitzen, die im SunPlex-System eingesetzt wird.

Typografische Konventionen

Die folgende Tabelle beschreibt die in diesem Buch verwendeten Schriftänderungen.

TABELLE P–1 Typografische Konventionen

Schriftart oder Symbol	Bedeutung	Beispiel
`AaBbCc123`	Die Namen aller Befehle, Dateien und Verzeichnisse; Bildschirmausgabe des Computers	Bearbeiten Sie Ihre `.login`-Datei. Verwenden Sie `ls -a` , um eine Liste aller Dateien zu erhalten. `Rechnername% Sie haben eine neue Nachricht.`
`AaBbCc123`	Die Eingaben des Benutzers, im Gegensatz zu den Bildschirmausgaben des Computers	`Rechner_name%` **`su`** `Passwort:`
AaBbCc123	Befehlszeilen-Variable: durch einen realen Namen oder Wert ersetzen	Um eine Datei zu löschen, geben Sie Folgendes ein: **`rm`** *Dateiname*.
AaBbCc123	Buchtitel, neue Wörter oder Begriffe bzw. hervorzuhebende Wörter.	Lesen Sie dazu auch Kapitel 6 im *Benutzerhandbuch*. Diese werden *class*-Optionen genannt. Sie *müssen* als root angemeldet sein, um dies zu tun.

Beispiele zu Shell-Eingabeaufforderungen in Befehlen

Die folgende Tabelle zeigt die Standard-Systemeingabeaufforderung und die Superbenutzer-Eingabeaufforderung für die C-Shell, die Bourne-Shell und die Korn-Shell.

TABELLE P–2 Shell-Eingabeaufforderungen

Shell	Eingabeaufforderung
C Shell-Eingabeaufforderung	`Rechnername%`
C Shell-Superbenutzer-Eingabeaufforderung	`Rechnername#`

TABELLE P–2 Shell-Eingabeaufforderungen *(Fortsetzung)*

Shell	Eingabeaufforderung
Bourne Shell- und Korn Shell-Eingabeaufforderung	`$`
Bourne Shell- und Korn Shell-Superbenutzer-Eingabeaufforderung	#

Verwandte Dokumentation

Thema	Titel	Teilenummer
Installation	*Sun Cluster 3.1 10/03 Handbuch Softwareinstallation*	817-4252-10
Hardware	*Sun Cluster 3.x Hardware Administration Manual*	817-0168
	Sun Cluster 3.x Hardware Administration Collection unter `http://docs.sun.com/db/coll/1024.1/`	
Datendienste	*Sun Cluster 3.1 Data Service Planning and Administration Guide*	817–3305
	Sun Cluster 3.1 Data Services 10/03 Collection unter `http://docs.sun.com/db/coll/573.11`	
API-Entwicklung	*Sun Cluster 3.1 10/03 Entwicklerhandbuch Datendienste*	817-4264-10
Verwaltung	*Sun Cluster 3.1 10/03 Handbuch Systemverwaltung*	817-4246-10
Fehlermeldungen	*Sun Cluster 3.1 10/03 Error Messages Guide*	817–0521
Online-Dokumentation	*Sun Cluster 3.1 10/03 Reference Manual*	817–0522
Versionshinweise	*Sun Cluster 3.1 10/03 Versionshinweise*	817-4854-10
	Sun Cluster 3.x Release Notes Supplement	816-3381

Zugriff auf die Online-Dokumentation von Sun

Über die Website docs.sun.com^{SM} erhalten Sie Zugriff auf die technische Online-Dokumentation von Sun. Sie können das Archiv unter docs.sun.com durchsuchen oder nach einem bestimmten Buchtitel oder Thema suchen. Die URL lautet: `http://docs.sun.com`.

Bestellen von Sun-Dokumentation

Ausgewählte Produktdokumentationen bietet Sun Microsystems auch in gedruckter Form an. Eine Liste dieser Dokumente und Hinweise zum Bezug finden Sie unter „Buy printed documentation" auf der Website http://docs.sun.com.

Hilfe anfordern

Wenden Sie sich im Falle von Problemen bei Installation oder Verwendung des SunPlex-Systems an Ihren Dienstleister, und geben Sie folgende Informationen an:

- Ihren Namen und E-Mail-Adresse (ggf.)
- Firmennamen, Adresse, Telefonnummer
- Modell und Seriennummern des Systems
- Versionsnummer des Betriebssystems (z.B. Solaris 9)
- Versionsnummer der Sun Cluster-Software (z.B. Sun Cluster 3.1)

Sammeln Sie für Ihren Dienstleister mithilfe folgender Befehle Informationen zu allen Knoten.

Befehl	Funktion
`prtconf -v`	Zeigt die Größe des Systemspeichers an und gibt Informationen zu Peripheriegeräten zurück.
`psrinfo -v`	Zeigt Informationen zu Prozessoren an.

Befehl	Funktion
`showrev -p`	Gibt die installierten Korrekturversionen zurück.
`prtdiag -v`	Zeigt Informationen zu Systemdiagnosen an.
`scinstall -pv`	Zeigt Informationen zu Sun Cluster-Software- und Packetversion an.
`scstat`	Liefert einen Schnappschuss des Cluster-Status.
`scconf -p`	Listet die Cluster-Konfigurationsinformationen auf.
`scrgadm -p`	Zeigt Informationen zu installierten Ressourcen, Ressourcengruppen und Ressourcentypen an.

Halten Sie zudem den Inhalt der Datei `/var/adm/messages` bereit.

KAPITEL 1

Einführung und Überblick

Das SunPlex-System ist eine integrierte Lösung aus Hardware und Sun Cluster-Software zur Erstellung hoch verfügbarer und von Scalable-Diensten.

Das Sun Cluster 3.1 10/03 Konzepthandbuch liefert die benötigten konzeptionellen Informationen für die wichtigste Zielgruppe der SunPlex -Dokumentation. Zu dieser Zielgruppe gehören

- Dienstleister, die Cluster-Hardware installieren und warten,
- Systemverwalter, die Sun Cluster-Software installieren, konfigurieren und verwalten,
- Anwendungsentwickler, welche die derzeit noch nicht in Sun Cluster enthaltenen Failover- und Scalable-Dienste entwickeln.

Dieses Buch vermittelt zusammen mit der restlichen SunPlex-Dokumentationsreihe einen vollständigen Überblick über das SunPlex-System.

Kapitelinhalt

- Einführung in das SunPlex-System und Übersicht auf höchster Ebene,
- Beschreibung der unterschiedlichen Standpunkte der jeweiligen SunPlex-Zielgruppen,
- Identifizierung der Schlüsselkonzepte, deren Verständnis für das Arbeiten mit dem SunPlex-System notwendig ist,
- Zuordnung der Schlüsselkonzepte zur SunPlex-Dokumentation, die Verfahren und verwandte Informationen enthält,
- Zuordnung Cluster-spezifischer Aufgaben zur Dokumentation, in der die Verfahren zur Ausführung dieser Aufgaben enthalten sind.

Einführung in das SunPlex-System

Das SunPlex-System erweitert die Solaris-Betriebsumgebung auf ein Cluster-Betriebssystem. Ein Cluster oder Plex ist eine Reihe von locker gekoppelten Computerknoten, die sich für den Client als ein einziges Netzwerk bzw. eine einzige Anwendung mit Datenbanken, Webdiensten und Dateidiensten darstellt.

Jeder Cluster-Knoten ist ein Standalone-Server, der seine eigenen Prozesse ausführt. Diese Prozesse kommunizieren miteinander und treten (für einen Netzwerk-Client) als ein einziges System auf. Im Zusammenspiel stellen sie Benutzern Anwendungen, Systemressourcen und Daten zur Verfügung.

Ein Cluster bietet im Vergleich zu traditionellen Einzelserversystemen mehrere Vorteile. Zu diesen Vorteilen gehört die Unterstützung von Failover- und Scalable-Diensten, die Fähigkeit zum modularen Wachsen und ein niedriger Anschaffungspreis im Vergleich zu traditionellen fehlertoleranten Hardwaresystemen.

Ziele des SunPlex-Systems

- Reduzieren oder Beseitigen der Systemausfallzeiten aufgrund von Softwarefehlern oder Hardwareausfällen.
- Sicherstellen der Verfügbarkeit von Daten und Anwendungen für Endbenutzer, unabhängig von der Art des Fehlers, der normalerweise zum Ausfall eines Einzelserversystems führt.
- Erhöhen der Durchsatzleistung von Anwendungen, indem Dienste durch die Aufnahme weiterer Knoten in den Cluster auf zusätzliche Prozessoren skalierbar gemacht werden.
- Verbessern der Systemverfügbarkeit, weil Sie das System pflegen können, ohne den ganzen Cluster herunterfahren zu müssen.

Hochverfügbarkeit versus Fehlertoleranz

Das SunPlex-System ist als *hoch verfügbares* System (HA-System) ausgelegt, das heißt als ein System, das einen nahezu kontinuierlichen Zugriff auf Daten und Anwendungen bietet.

Im Gegensatz dazu sorgen *fehlertolerante* Hardwaresysteme für einen kontinuierlichen Zugriff auf Daten und Anwendungen, verursachen aber aufgrund der spezialisierten Hardware höhere Kosten. Zudem überbrücken fehlertolerante Systeme normalerweise keine Softwarefehler.

Das SunPlex-System setzt eine Kombination aus Hardware und Software ein, um Hochverfügbarkeit zu erzielen. Redundante Cluster-Interconnects, Speicherung und öffentliche Netzwerke schützen vor kritischen Ausfallpunkten (Single Points of

Failure, SPOF). Die Cluster-Software überwacht kontinuierlich den Zustand der Mitgliedsknoten und hindert fehlerhaft arbeitende Knoten an der Teilnahme im Cluster, um die Daten vor Schäden zu schützen. Der Cluster überwacht auch Dienste und die abhängigen Systemressourcen und sorgt bei Ausfällen für ein Failover oder einen Neustart der Dienste.

Fragen und Antworten zur Hochverfügbarkeit finden Sie unter „Häufig gestellte Fragen zur Hochverfügbarkeit“ auf Seite 91.

Failover und Skalierbarkeit im SunPlex-System

Mit dem SunPlex-System können Sie entweder *Failover-* oder *Scalable*-Dienste implementieren. Im Allgemeinen sorgt ein Failover-Dienst nur für Hochverfügbarkeit (Redundanz), während ein Scalable-Dienst Hochverfügbarkeit mit einer besseren Leistung verbindet. Ein einzelner Cluster kann sowohl Failover- als auch Scalable-Dienste unterstützen.

Failover-Dienste

Failover ist der Vorgang, bei dem der Cluster automatisch einen Dienst von einem ausgefallenen Primärknoten auf einen dafür vorgesehenen Sekundärknoten verschiebt. Mit Failover sorgt die Sun Cluster-Software für Hochverfügbarkeit.

Bei einem Failover nimmt der Client möglicherweise eine kurze Unterbrechung im Dienst wahr und muss ggf. die Verbindung neu herstellen, nachdem das Failover abgeschlossen ist. Die Clients nehmen jedoch den realen Server, der den Dienst zur Verfügung stellt, nicht wahr.

Scalable-Dienste

Während es beim Failover um Redundanz geht, sorgt die Skalierbarkeit für eine konstante Antwortzeit oder einen von der Last unabhängigen konstanten Datendurchsatz. Ein Scalable-Dienst ermöglicht die wirksame Zusammenarbeit der vielen Knoten eines Clusters bei der gemeinsamen Ausführung einer Anwendung und sorgt damit für eine bessere Leistung. In einer Scalable-Konfiguration kann jeder Cluster-Knoten Daten liefern und Client-Abfragen verarbeiten.

Spezifischere Informationen zu Failover- und Scalable-Diensten finden Sie unter „Datendienste“ auf Seite 59.

Drei Aspekte des SunPlex-Systems

In diesem Abschnitt werden drei unterschiedliche Aspekte des SunPlex-Systems sowie die Schlüsselkonzepte und die Dokumentation beschrieben, die für den jeweiligen Aspekt wichtig sind. Diese drei Aspekte sind vertreten durch:

- Hardware-Installations- und Servicepersonal
- Systemverwalter
- Anwendungsprogrammierer

Hardwareinstallation und -service

Für Hardwarespezialisten stellt sich das SunPlex-System als eine Sammlung aus Standard-Hardware dar, zu der Server, Netzwerke und Speicher gehören. Diese Komponenten werden mit Kabeln verbunden, so dass jede Komponente eine Sicherung hat und kein Single Point of Failure (Ausfallpunkt) mehr gegeben ist.

Schlüsselkonzepte – Hardware

Hardwarespezialisten müssen die folgenden Cluster-Konzepte verstehen.

- Cluster-Hardwarekonfiguration und Verkabelung
- Installation und Wartung (Hinzufügen, Entfernen, Ersetzen):
  - Netzwerkschnittstellen-Komponenten (Adapter, Verbindungspunkte, Kabel)
  - Schnittstellenkarten für Platten
  - Platten-Arrays
  - Plattenlaufwerke
  - Verwaltungskonsole und Konsolenzugriffsgerät
- Konfiguration der Verwaltungskonsole und des Konsolenzugriffsgeräts

Empfohlene Verweise auf Hardwarekonzepte

Die folgenden Abschnitte enthalten relevante Informationen zu den vorstehenden Schlüsselkonzepten:

- „Cluster-Knoten “ auf Seite 23
- „Multihostplatten“ auf Seite 25
- „Lokale Platten“ auf Seite 26
- „Cluster-Interconnect“ auf Seite 27
- „Öffentliche Netzwerkschnittstellen“ auf Seite 28
- „Client-Systeme“ auf Seite 28

- „Verwaltungskonsole“ auf Seite 29
- „Konsolenzugriffsgerät“ auf Seite 28
- „Geclusterte-Paare-Topologie“ auf Seite 30
- „N+1- (Stern)-Topologie“ auf Seite 32

Relevante SunPlex-Dokumentation

Folgendes SunPlex-Dokument enthält Verfahren und Informationen zu Konzepten für den Hardwareservice:

- *Sun Cluster 3.1 Hardware Collection*

Systemverwalter

Für den Systemverwalter stellt sich das SunPlex-System als Satz aus verkabelten Servern (Knoten) und gemeinsam genutzten Speichergeräten dar. Der Systemverwalter sieht Folgendes:

- In die Solaris-Software integrierte Cluster-Spezialsoftware zum Überwachen der Konnektivität zwischen den Cluster-Knoten,
- Spezialsoftware zum Überwachen des Zustands der auf den Cluster-Knoten laufenden Benutzer-Anwendungsprogramme,
- Datenträgerverwaltungs-Software zum Konfigurieren und Verwalten von Platten,
- Cluster-Spezialsoftware, die allen Knoten Zugriff auf alle Speichergeräte ermöglicht, auch auf solche, die nicht direkt mit Platten verbunden sind,
- Cluster-Spezialsoftware, dank der Dateien auf allen Knoten angezeigt werden, als seien sie lokal mit diesem Knoten verbunden.

Schlüsselkonzepte – Systemverwaltung

Systemverwalter müssen die folgenden Konzepte und Prozesse verstehen:

- Interaktion zwischen Hardware- und Softwarekomponenten
- Allgemeiner Ablauf der Cluster-Installation und -Konfiguration, einschließlich:
  - Installieren der Solaris-Betriebsumgebung
  - Installieren und Konfigurieren der Sun Cluster-Software
  - Installieren und Konfigurieren eines Datenträger-Managers
  - Installieren und Konfigurieren der Anwendungssoftware für den Cluster-Betrieb
  - Installieren und Konfigurieren der Sun Cluster-Datendienstsoftware
- Cluster-Verwaltungsverfahren für das Hinzufügen, Entfernen, Ersetzen und Pflegen der Cluster-Hardware- und Softwarekomponenten

- Konfigurationsänderungen zur Leistungssteigerung

Empfohlene Verweise auf Konzepte für Systemverwalter

Die folgenden Abschnitte enthalten relevante Informationen zu den vorstehenden Schlüsselkonzepten:

- „Verwaltungsschnittstellen“ auf Seite 36
- „Hoch verfügbares Framework“ auf Seite 38
- „Globale Geräte“ auf Seite 41
- „Plattengerätegruppen“ auf Seite 42
- „Globaler Namensraum“ auf Seite 46
- „Cluster-Dateisysteme“ auf Seite 47
- „Quorum und Quorum-Geräte“ auf Seite 54
- „Datenträger-Manager“ auf Seite 59
- „Datendienste“ auf Seite 59
- „Ressourcen, Ressourcengruppen und Ressourcentypen“ auf Seite 72
- „Öffentliche Netzwerkadapter und IP Network Multipathing“ auf Seite 85
- Kapitel 4

Relevante SunPlex-Dokumentation – Systemverwalter

Die folgenden SunPlex-Dokumente enthalten Verfahren und Informationen zu Konzepten der Systemverwaltung:

- *Sun Cluster 3.1 Handbuch Softwareinstallation*
- *Sun Cluster 3.1 Handbuch Systemverwaltung*
- *Sun Cluster 3.1 Error Messages Guide*
- *Sun Cluster 3.1 Versionshinweise*
- *Sun Cluster 3.1 Release Notes Supplement*

Anwendungsprogrammierer

Das SunPlex-System stellt *Datendienste* für Anwendungen wie Oracle, NFS, DNS, Sun™ ONE Web Server, Apache Web Server und Sun™ ONE Directory Server zur Verfügung. Datendienste werden erstellt, indem Standard-Anwendungen für die Steuerung durch die Sun Cluster-Software konfiguriert werden. Die Sun Cluster-Software stellt Konfigurationsdateien und Verwaltungsmethoden zum Starten, Stoppen und Überwachen der Anwendungen zur Verfügung. Wenn Sie einen neuen Failover- oder Scalable-Dienst erstellen müssen, können Sie mit der Anwendungsprogrammierungsschnittstelle (API) von SunPlex und der DSET-API (API der Datendienst-Grundlagentechnologie) die Konfigurationsdateien und Verwaltungsmethoden erstellen, die zur Ausführung der Anwendung als Datendienst auf dem Cluster erforderlich sind.

Schlüsselkonzepte– Anwendungsprogrammierer

Anwendungsprogrammierer müssen Folgendes verstehen:

- Die Eigenschaften der Anwendung, um festzustellen, ob diese zur Ausführung als Failover- oder Scalable-Dienst eingerichtet werden kann.
- Die Sun Cluster-API, DSET-API sowie den "generischen" Datendienst. Die Programmierer müssen entscheiden, welches Tool sich am besten für das Schreiben von Programmen oder Skripts eignet, mit denen ihre Anwendung für die Cluster-Umgebung konfiguriert werden soll.

Empfohlene Verweise auf Konzepte für Anwendungsprogrammierer

Die folgenden Abschnitte enthalten relevante Informationen zu den vorstehenden Schlüsselkonzepten:

- „Datendienste“ auf Seite 59
- „Ressourcen, Ressourcengruppen und Ressourcentypen“ auf Seite 72
- Kapitel 4

Relevante SunPlex-Dokumentation – Anwendungsprogrammierer

Die folgenden SunPlex-Dokumente enthalten Verfahren und Informationen zu Konzepten für Anwendungsprogrammierer:

- *Sun Cluster 3.1 Entwicklerhandbuch Datendienste*
- *Sun Cluster 3.1 Data Services Installation and Configuration Guide*

SunPlex-Systemaufgaben

Alle SunPlex-Systemaufgaben erfordern ein gewisses konzeptionelles Vorverständnis. Die nachstehende Tabelle gibt einen Überblick auf höchster Ebene über die Aufgaben und die Dokumentation, in der die Aufgabenschritte beschrieben sind. Die den Konzepten gewidmeten Abschnitte in diesem Buch beschreiben, wie die Konzepte den Aufgaben zugeordnet sind.

TABELLE 1–1 Aufgabenzuordnung: Zuordnen von Benutzeraufgaben zur Dokumentation

Für diese Aufgabe...	Verwenden Sie diese Dokumentation...
Installieren von Cluster-Hardware	*Sun Cluster 3.1 Hardware Collection*
Installieren der Solaris-Software auf dem Cluster	*Sun Cluster 3.1 Handbuch Softwareinstallation*
Installieren der Sun™ Management Center-Software	*Sun Cluster 3.1 Handbuch Softwareinstallation*
Installieren und Konfigurieren der Sun Cluster-Software	*Sun Cluster 3.1 Handbuch Softwareinstallation*
Installieren und Konfigurieren der Datenträgerverwaltungs-Software	*Sun Cluster 3.1 Handbuch Softwareinstallation* Ihre Dokumentation zur Datenträgerverwaltung
Installieren und Konfigurieren von Sun Cluster-Datendiensten	*Sun Cluster 3.1 Data Services Installation and Configuration Guide*
Warten der Cluster-Hardware	*Sun Cluster 3.1 Hardware Collection*
Verwalten der Sun Cluster-Software	*Sun Cluster 3.1 Handbuch Systemverwaltung*
Verwalten der Datenträgerverwaltungs-Software	*Sun Cluster 3.1 Handbuch Systemverwaltung* und Ihre Dokumentation zur Datenträgerverwaltung
Verwalten von Anwendungssoftware	Ihre Anwendungsdokumentation
Problemidentifizierung und empfohlene Benutzeraktionen	*Sun Cluster 3.1 Error Messages Guide*
Erstellen eines neuen Datendienstes	*Sun Cluster 3.1 Entwicklerhandbuch Datendienste*

KAPITEL 2

Schlüsselkonzepte – Hardwaredienstleister

In diesem Kapitel werden die Schlüsselkonzepte im Zusammenhang mit den Hardwarekomponenten einer SunPlex-Systemkonfiguration beschrieben. Folgende Themen werden behandelt:

SunPlex-System-Hardwarekomponenten

Diese Informationen richten sich in erster Linie an Hardwaredienstleister. Diese Konzepte helfen Dienstleistern dabei, die Beziehungen zwischen den Hardwarekomponenten zu verstehen, bevor sie Cluster-Hardware installieren, konfigurieren oder warten. Auch für Cluster-Systemverwalter können diese Informationen als Hintergrund für das Installieren, Konfigurieren und Verwalten von Cluster-Software nützlich sein.

Ein Cluster besteht aus mehreren Hardwarekomponenten einschließlich:

- Cluster-Knoten mit lokalen Platten (nicht gemeinsam genutzt)
- Multihost-Speicher (von Knoten gemeinsam benutzte Platten)
- Wechselmedien (Bänder und CD-ROM)
- Cluster-Interconnect

- Öffentliche Netzwerkschnittstellen
- Client-Systeme
- Verwaltungskonsole
- Konsolenzugriffsgeräte

Mit dem SunPlex-System können Sie diese Komponenten in unterschiedlichen Konfigurationen kombinieren, die unter „Sun Cluster-Topologiebeispiele“ auf Seite 30 beschrieben sind.

Die nachstehende Abbildung zeigt ein Beispiel einer Cluster-Konfiguration.

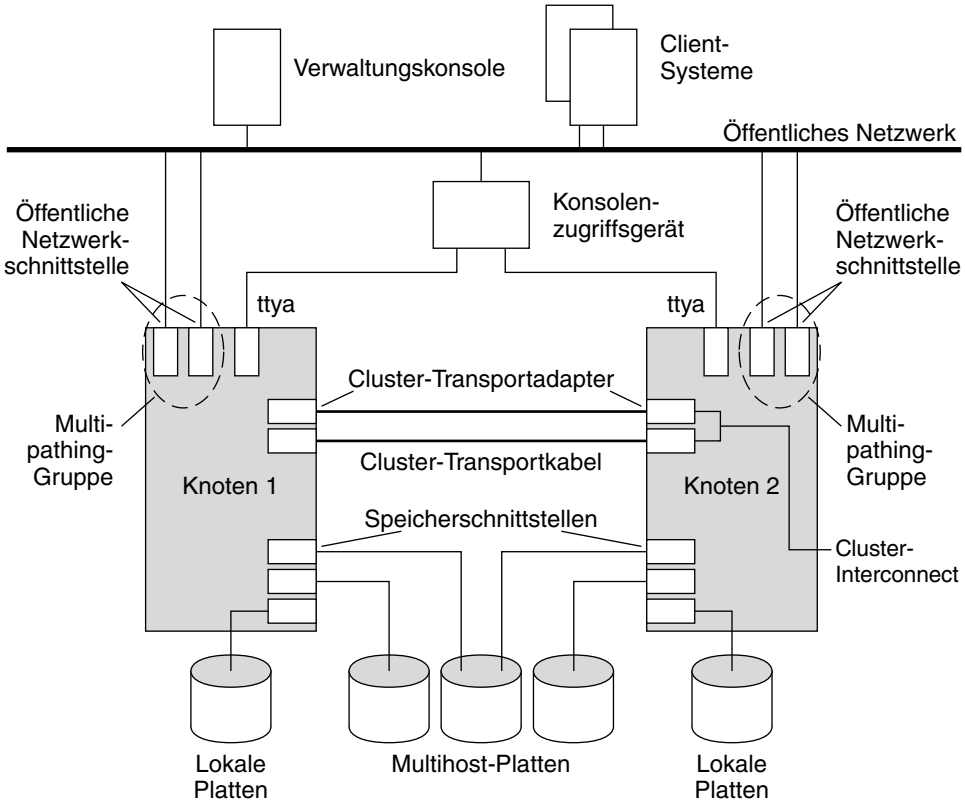


ABBILDUNG 2–1 Beispiel für eine Zwei-Knoten-Cluster-Konfiguration

Cluster-Knoten

Ein Cluster-Knoten ist ein Rechner, auf dem sowohl die Solaris-Betriebsumgebung als auch die Sun Cluster-Software läuft und der entweder ein aktuelles (ein *Cluster-Mitglied*) oder ein potenzielles Mitglied des Clusters ist. Mit der Sun Cluster-Software können Sie von zwei bis acht Knoten in einen Cluster einbinden. Die unterstützten Knotenkonfigurationen finden Sie unter „Sun Cluster-Topologiebeispiele“ auf Seite 30.

Cluster-Knoten sind im allgemeinen an mindestens eine Multihostplatte angeschlossen. Nicht an Multihostplatten angeschlossene Knoten verwenden das Cluster-Dateisystem zum Zugriff auf die Multihostplatten. In einer Scalable-Dienste-Konfiguration können die Knoten beispielsweise Anfragen abwickeln, ohne direkt mit den Multihostplatten verbunden zu sein.

Daneben teilen Knoten in Konfigurationen paralleler Datenbanken den gleichzeitigen Zugriff auf alle Platten. Weitere Informationen zu Konfigurationen paralleler Datenbanken finden Sie unter „Multihostplatten“ auf Seite 25 und Kapitel 3.

Alle Knoten im Cluster sind unter einem gemeinsamen Namen zusammengefasst — dem Cluster-Namen —, der für den Cluster-Zugriff und die Cluster-Verwaltung verwendet wird.

Öffentliche Netzwerkadapter verbinden die Knoten mit den öffentlichen Netzwerken und ermöglichen den Zugriff der Clients auf den Cluster.

Cluster-Mitglieder kommunizieren über mindestens ein real unabhängiges Netzwerk miteinander. Dieser Satz aus real unabhängigen Netzwerken wird als *Cluster-Interconnect* bezeichnet.

Jeder Knoten im Cluster nimmt wahr, wenn ein anderer Knoten dem Cluster beitritt oder diesen verlässt. Daneben nimmt jeder Knoten sowohl die lokalen Ressourcen als auch die Ressourcen auf den anderen Cluster-Knoten wahr.

Knoten innerhalb desselben Clusters sollten eine vergleichbare Verarbeitungs-, Speicher- und E/A-Kapazität aufweisen, so dass ein Failover ohne nennenswerten Leistungsabfall möglich ist. Aufgrund eines möglichen Failovers muss jeder Knoten über genügend Überkapazität verfügen, um die Arbeitslast aller Knoten zu übernehmen, für die sie als Sicherung oder Sekundärknoten fungieren.

Jeder Knoten bootet sein eigenes, individuelles Root-Dateisystem (/).

Softwarekomponenten für Cluster-Hardware-Mitglieder

Folgende Software muss installiert sein, um als Cluster-Mitglied funktionieren zu können:

- Solaris-Betriebsumgebung
- Sun Cluster-Software

- Datendienstanwendung
- Datenträgerverwaltung (Solaris Volume Manager™ oder VERITAS Volume Manager)

  Davon ausgenommen sind Konfigurationen, die Hardware-RAID (Hardware Redundant Array of Independent Disks) verwenden. Diese Konfiguration benötigt ggf. keinen Software-Datenträger-Manager wie Solaris Volume Manager oder VERITAS Volume Manager.

Im *Sun Cluster 3.1 Handbuch Softwareinstallation* finden Sie Informationen zur Installation der Solaris-Betriebsumgebung, des Sun Clusters und der Datenträgerverwaltungs-Software.

Informationen zur Installation und Konfiguration von Datendiensten finden Sie im *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Konzeptionelle Informationen zu den vorgenannten Softwarekomponenten finden Sie in Kapitel 3.

Die nachstehende Abbildung bietet einen Überblick auf höchster Ebene über die Softwarekomponenten, die zusammen die Sun Cluster-Softwareumgebung bilden.

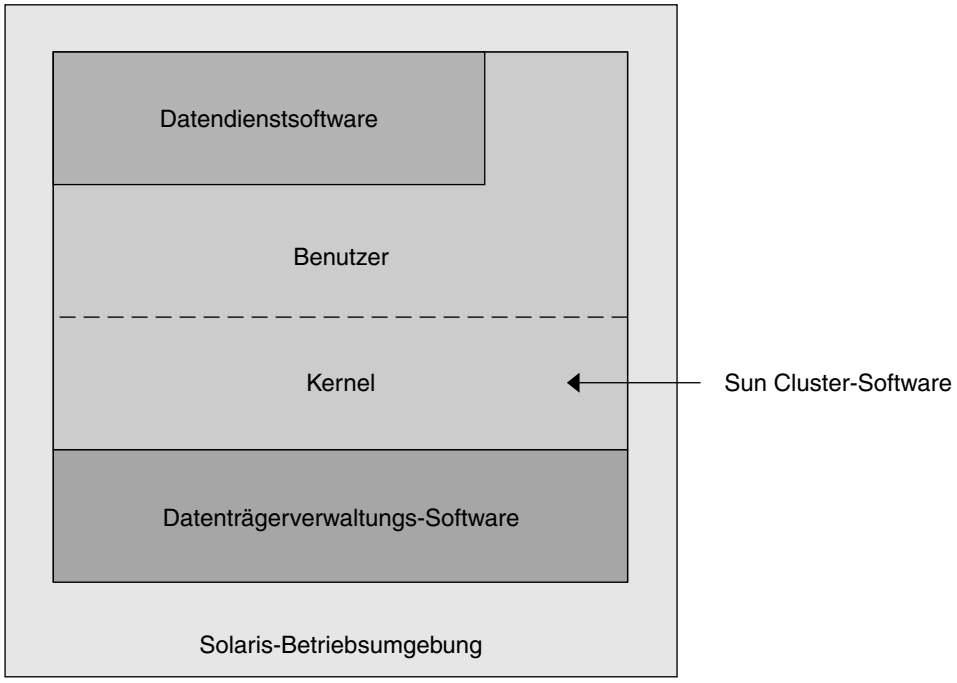


ABBILDUNG 2–2 Die Beziehung der Sun Cluster-Softwarekomponenten auf höchster Ebene

Fragen und Antworten zu Cluster-Mitgliedern finden Sie in Kapitel 4.

Multihostplatten

Platten, die mit mehr als einem Knoten gleichzeitig verbunden werden können, sind Multihostplatten. In einer Sun Cluster-Umgebung sorgt die Multihostspeicherung für hoch verfügbare Platten. Sun Cluster benötigt Multihostspeicher für Zwei-Knoten-Cluster zur Einrichtung des Quorums. Cluster mit mehr als drei Knoten benötigen keinen Multihostspeicher.

Multihostplatten haben folgende Eigenschaften.

- Sie tolerieren das Versagen einzelner Knoten.
- Sie speichern Anwendungsdaten und können auch binäre Anwendungsdateien und Konfigurationsdateien speichern.
- Sie schützen vor Knotenversagen. Wenn Client-Anfragen über einen Knoten auf Daten zugreifen und dieser Knoten ausfällt, werden die Anfragen umgeschaltet und über einen anderen Knoten abgewickelt, der mit denselben Platten direkt verbunden ist.
- Der Zugriff erfolgt entweder global über einen Primärknoten, der die Platten "unterstützt", oder durch einen direkten gleichzeitigen Zugriff über lokale Pfade. Die derzeit einzige Anwendung, die mit direkten gleichzeitigen Zugriffen arbeitet, ist Oracle Parallel Server.

Ein Datenträger-Manager sorgt bei gespiegelten oder RAID-5-Konfigurationen für die Datenredundanz der Multihostplatten. Derzeit unterstützt Sun Cluster die Datenträger-Manager Solaris Volume Manager™ und VERITAS Volume Manager sowie den RDAC RAID-5-Hardware-Controller auf verschiedenen RAID-Hardwareplattformen.

Die Kombination aus Multihostplatten und Plattenspiegelung sowie Striping schützt vor Knoten- und Einzelplattenversagen.

Fragen und Antworten zum Multihostspeicher finden Sie in Kapitel 4.

SCSI-Multi-Initiator

Dieser Abschnitt betrifft nur SCSI-Speichergeräte und keine Glasfaserkanal-Speicher, die für Multihostplatten eingesetzt werden.

Auf einem Standalone-Server steuert der Serverknoten die SCSI-Busaktivitäten durch die SCSI-Host-Adapterschaltung, die diesen Server mit einem bestimmten SCSI-Bus verbindet. Diese SCSI-Host-Adapterschaltung wird als *SCSI-Initiator* bezeichnet. Die Schaltung initiiert alle Busaktivitäten für diesen SCSI-Bus. Die SCSI-Standardadresse der SCSI-Host-Adapter in Sun-Systemen ist 7.

Cluster-Konfigurationen teilen ihren Speicher unter Verwendung von Multihostplatten auf mehrere Serverknoten auf. Wenn der Cluster-Speicher aus symmetrischen oder unsymmetrischen SCSI-Geräten besteht, wird die Konfiguration als Multi-Initiator-SCSI bezeichnet. Wie aus dieser Terminologie hervorgeht, ist mehr als ein SCSI-Initiator auf dem SCSI-Bus vorhanden ist.

Die SCSI-Spezifikation erfordert, dass jedes Gerät im SCSI-Bus eine einmalige SCSI-Adresse hat. (Der Host-Adapter ist ebenfalls ein Gerät im SCSI-Bus.) Die Standard-Hardwarekonfiguration führt in einer Multi-Initiator-Umgebung zu einem Konflikt, weil alle SCSI-Host-Adapter standardmäßig auf 7 gesetzt werden.

Lassen Sie zur Beseitigung dieses Konflikts auf jedem SCSI-Bus einen der SCSI-Host-Adapter mit der SCSI-Adresse 7, und stellen Sie für die anderen Host-Adapter ungenutzte SCSI-Adressen ein. Eine zweckmäßige Planung schreibt vor, dass sowohl die derzeit ungenutzten als auch die potenziell ungenutzten Adressen zu den "ungenutzten" SCSI-Adressen gehören. Ein Beispiel für zukünftige, ungenutzte Adressen ist das Aufstocken des Speichers durch das Installieren neuer Laufwerke an freien Laufwerk-Steckplätzen. In den meisten Konfigurationen ist 6 die verfügbare SCSI-Adresse für einen zweiten Host-Adapter.

Sie können die ausgewählten SCSI-Adressen für diese Host-Adapter mit der Open Boot PROM (OBP)-Eigenschaft `scsi-initiator-id` einstellen. Sie können diese Eigenschaft global für einen Knoten oder für jeden Host-Adapter einzeln einstellen. Anweisungen zum Einstellen einer einmaligen `scsi-initiator-id`-Eigenschaft für jeden SCSI-Host-Adapter finden Sie im jeweiligen Kapitel für jedes Plattengehäuse im *Sun Cluster 3.1 Hardware Collection*.

Lokale Platten

Lokale Platten sind Platten, die nur mit einem einzigen Knoten verbunden sind. Sie sind aus diesem Grund nicht vor Knotenversagen geschützt (nicht hoch verfügbar). Dennoch sind alle Platten, auch lokale Platten, im globalen Namensraum enthalten und als *globale Geräte* konfiguriert. Aus diesem Grund sind die Platten selbst von allen Cluster-Knoten aus sichtbar.

Sie können die Dateisysteme auf lokalen Platten für andere Knoten verfügbar machen, indem Sie diese unter einen globalen Einhängepunkt stellen. Versagt der Knoten, an dem derzeit eines dieser globalen Dateisysteme eingehängt ist, kann kein Knoten auf dieses Dateisystem zugreifen. Mithilfe eines Datenträger-Managers können Sie diese Platten spiegeln, so dass ein Fehler nicht dazu führt, dass kein Zugriff mehr auf diese Dateisysteme besteht; Datenträger-Manager schützen aber nicht vor Knotenversagen.

Weitere Informationen zu globalen Geräten finden Sie im Abschnitt „Globale Geräte" auf Seite 41.

Wechselmedien

Wechselmedien wie Band- und CD-ROM-Laufwerke werden in einem Cluster unterstützt. Im Allgemeinen installieren, konfigurieren und warten Sie diese Geräte ebenso wie in einer Nicht-Cluster-Umgebung. Diese Geräte werden in Sun Cluster als globale Geräte konfiguriert, so dass jeder Knoten im Cluster auf jedes Gerät zugreifen kann. Informationen zum Installieren und Konfigurieren von Wechselmedien finden Sie unter *Sun Cluster 3.1 Hardware Collection*.

Weitere Informationen zu globalen Geräten finden Sie im Abschnitt „Globale Geräte“ auf Seite 41.

Cluster-Interconnect

Der *Cluster-Interconnect* ist die reale Konfiguration der Geräte, die für die Übertragung von Cluster-privaten Kommunikationen und Datendienstkommunikationen zwischen Cluster-Knoten eingesetzt wird. Der Interconnect wird umfassend für Cluster-interne Kommunikationen verwendet und kann dadurch die Leistung begrenzen.

Nur Cluster-Knoten können mit dem Cluster-Interconnect verbunden werden. Das Sun Cluster-Sicherheitsmodell setzt voraus, dass nur Cluster-Knoten real Zugriff auf den Cluster-Interconnect haben.

Alle Knoten müssen mit dem Cluster-Interconnect über mindestens zwei redundante, real unabhängige Netzwerke oder Pfade verbunden sein, um einen Single Point of Failure (Ausfallpunkt) zu vermeiden. Sie können mehrere real unabhängige Netzwerke (bis zu sechs) zwischen zwei beliebigen Knoten einrichten. Der Cluster-Interconnect besteht aus drei Hardwarekomponenten: Adapter, Verbindungspunkten und Kabeln.

Die nachstehende Liste beschreibt jede dieser Hardwarekomponenten.

- Adapter – Die Netzwerk-Schnittstellenkarten, die sich in jedem Cluster-Knoten befinden. Ihre Namen bestehen aus dem Gerätenamen, unmittelbar gefolgt von der realen Einheitsnummer, zum Beispiel qfe2. Manche Adapter haben nur eine reale Netzwerkverbindung, andere hingegen, wie die qfe-Karte, haben mehrere reale Verbindungen. Manche enthalten auch sowohl Netzwerk- als auch Speicherschnittstellen.

  Ein Netzwerkadapter mit mehreren Schnittstellen kann zu einem Single Point of Failure werden, wenn der ganze Adapter ausfällt. Für maximale Verfügbarkeit sollten Sie Ihren Cluster so planen, dass der einzige Pfad zwischen zwei Knoten nicht von einem einzigen Netzwerkadapter abhängt.
- Verbindungspunkte – Die Schalter, die außerhalb der Cluster-Knoten liegen. Sie übernehmen Pass-Through- und Umschaltfunktionen, damit Sie mehr als zwei Knoten zusammenschließen können. In einem Zwei-Knoten-Cluster benötigen Sie keine Verbindungspunkte, weil die Knoten mit redundanten realen Kabeln, die an die redundanten Adapter auf jedem Knoten angeschlossen sind, direkt miteinander

verbunden werden können. Für Konfigurationen mit mehr als zwei Knoten sind normalerweise Verbindungspunkte erforderlich.

- Kabel – Die realen Verbindungen, die entweder zwischen zwei Netzwerkadaptern oder zwischen einem Adapter und einem Verbindungspunkt bestehen.

Fragen und Antworten zum Cluster-Interconnect finden Sie in Kapitel 4.

Öffentliche Netzwerkschnittstellen

Die Verbindung der Clients mit dem Cluster erfolgt über die öffentlichen Netzwerkschnittstellen. Jede Netzwerkadapterkarte kann an ein oder mehrere öffentliche Netzwerke angeschlossen sein, je nachdem, ob die Karte mehrere Hardwareschnittstellen hat. Sie können Knoten mit mehreren öffentlichen Netzwerkschnittstellenkarten einrichten, die so konfiguriert sind, dass mehrere Karten aktiv sind und dadurch gegenseitig als Failover-Sicherung dienen. Wenn einer der Adapter ausfällt, wird die IP Network Multipathing-Software aufgerufen, um ein Failover von der fehlerhaften Schnittstelle auf einen anderen Adapter in der Gruppe auszuführen.

Es gibt keine hardwarespezifischen Erwägungen im Zusammenhang mit der Cluster-Bildung für öffentliche Netzwerke.

Fragen und Antworten zu öffentlichen Netzwerken finden Sie in Kapitel 4.

Client-Systeme

Client-Systeme beinhalten Workstations oder andere Server, die über das öffentliche Netzwerk auf den Cluster zugreifen. Clientseitige Programme verwenden Daten oder andere Dienste, die von auf dem Cluster laufenden serverseitigen Anwendungen zur Verfügung gestellt werden.

Client-Systeme sind nicht hoch verfügbar. Daten und Anwendungen auf dem Cluster sind hoch verfügbar.

Fragen und Antworten zu Client-Systemen finden Sie in Kapitel 4.

Konsolenzugriffsgerät

Sie benötigen Konsolenzugriff auf alle Cluster-Knoten. Verwenden Sie für den Konsolenzugriff den Terminal-Konzentrator, den Sie mit Ihrer Cluster-Hardware erworben haben, den System Service Processor (SSP) auf Sun Enterprise E10000™-Servern, den System-Controller auf Sun Fire™-Servern oder ein anderes Gerät, das auf `ttya` auf jedem Knoten zugreifen kann.

Bei Sun ist nur ein unterstützter Terminal-Konzentrator verfügbar, dessen Verwendung optional ist. Der Terminal-Konzentrator ermöglicht den Zugriff auf `/dev/console` auf jedem Knoten über ein TCP/IP-Netzwerk. Damit kann auf Konsolenebene von einer Remote-Workstation im Netzwerk aus auf jeden Knoten zugegriffen werden.

Der System Service Processor (SSP) stellt Konsolenzugriff für Sun Enterprise E10000-Server zur Verfügung. Der SSP ist ein Rechner in einem Ethernet-Netzwerk, der so konfiguriert wurde, dass er den Sun Enterprise E10000-Server unterstützt. Der SSP ist die Verwaltungskonsole für den Sun Enterprise E10000-Server. Mit der Sun Enterprise E10000-Netzwerkkonsolenfunktion kann jede Workstation im Netzwerk eine Host-Konsolensitzung öffnen.

Zu den weiteren Zugriffsmethoden auf die Konsole gehören andere Terminal-Konzentratoren, `tip(1)`-Zugriff über den seriellen Port von einem anderen Knoten aus und unintelligente Terminals. Sie können Tastaturen und Monitore von Sun™ oder andere Geräte für den seriellen Port verwenden, wenn Ihr Hardwaredienstleister diese unterstützt.

Verwaltungskonsole

Sie können eine dedizierte UltraSPARC ™-Workstation, die als *Verwaltungskonsole* bezeichnet wird, für die Verwaltung des aktiven Clusters verwenden. Normalerweise wird Verwaltungstool-Software wie der Cluster-Steuerbereich (Cluster Control Pane, CCP) und das Sun Cluster-Modul für das Sun Management Center™-Produkt in der Verwaltungskonsole installiert und ausgeführt. Mit `cconsole` im CCP können Sie mehr als eine Knotenkonsole gleichzeitig verbinden. Weitere Informationen zur Verwendung des CCP finden Sie im *Sun Cluster 3.1 Handbuch Systemverwaltung*.

Die Verwaltungskonsole ist kein Cluster-Knoten. Sie setzen die Verwaltungskonsole für den Remote-Zugriff auf Cluster-Knoten entweder über öffentliche Netzwerke oder optional über einen netzwerkbasierten Terminal-Konzentrator ein. Wenn Ihr Cluster aus der Sun Enterprise E10000-Plattform besteht, müssen Sie die Möglichkeit haben, sich von der Verwaltungskonsole aus beim System Service Processor (SSP) anzumelden und die Verbindung mit dem Befehl `netcon(1M)` herzustellen.

In der Regel konfigurieren Sie die Knoten ohne Monitore. Dann greifen Sie mit einer `telnet`-Sitzung von der Verwaltungskonsole aus, die mit einem Terminal-Konzentrator verbunden ist, auf die Konsole des Knotens zu, und vom Terminal-Konzentrator aus auf den seriellen Port des Knotens. (Bei einem Sun Enterprise E10000-Server stellen Sie die Verbindung über den System Service Processor her). Weitere Informationen finden Sie unter „Konsolenzugriffsgerät“ auf Seite 28.

Sun Cluster erfordert zwar keine dedizierte Verwaltungskonsole, aber diese bietet folgende Vorteile:

- Sie ermöglicht eine zentralisierte Cluster-Verwaltung durch das Gruppieren von Konsolen- und Verwaltungstools auf demselben Rechner.
- Sie sorgt für eine potenziell schnellere Problemlösung durch Ihren Hardware-Dienstleister.

Fragen und Antworten zur Verwaltungskonsole finden Sie in Kapitel 4.

Sun Cluster-Topologiebeispiele

Eine Topologie ist das Verbindungsschema, nach dem die Cluster-Knoten mit den im Cluster verwendeten Speicherplattformen verbunden sind. Sun Cluster unterstützt jede Topologie, die folgende Richtlinien einhält.

- Sun Cluster unterstützt maximal acht Knoten in einem Cluster, unabhängig von den Speicherkonfigurationen, die Sie implementieren.
- Ein gemeinsam genutztes Speichergerät kann mit so vielen Knoten verbunden werden, wie vom Speichergerät unterstützt werden.
- Gemeinsam genutzte Speichergeräte müssen nicht mit allen Knoten im Cluster verbunden sein. Diese Speichergeräte müssen jedoch mit mindestens zwei Knoten verbunden sein.

Sun Cluster schreibt für die Konfiguration eines Clusters keine spezifischen Topologien vor. Die Beschreibung der folgenden Topologien soll das Vokabular zur Erläuterung eines Cluster-Verbindungsschemas erschließen. Diese Topologien zeigen typische Verbindungsschemata.

- Geclusterte Paare
- Paar+N
- N+1 (Stern)
- N*N (Scalable)

Die folgenden Abschnitte enthalten Beispielgrafiken für jede Topologie.

Geclusterte-Paare-Topologie

Eine Geclusterte-Paare-Topologie besteht aus zwei oder mehr Knotenpaaren, die unter einem einzigen Cluster-Verwaltungs-Framework arbeiten. In dieser Konfiguration wird ein Failover nur zwischen einem Paar ausgeführt. Dabei sind jedoch alle Knoten über den Cluster-Interconnect verbunden und arbeiten unter der Sun Cluster-Softwaresteuerung. Mit dieser Topologie können Sie eine parallele Datenbankanwendung auf einem Paar und eine Failover- oder Scalable-Anwendung auf einem anderen Paar ausführen.

Mithilfe des Cluster-Dateisystems können Sie auch eine Konfiguration mit zwei Paaren erstellen, bei der mehr als zwei Knoten einen Scalable-Dienst oder eine parallele Datenbank ausführen, selbst wenn nicht alle Knoten mit den Platten verbunden sind, auf denen die Anwendungsdaten gespeichert werden.

Die nachstehende Abbildung zeigt eine Geclusterte-Paare-Konfiguration.

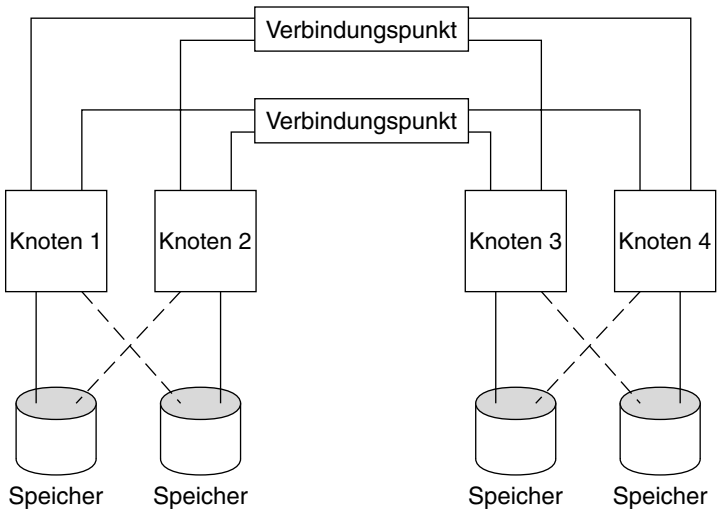


ABBILDUNG 2–3 Geclusterte-Paare-Topologie

Paar+N-Topologie

Die Paar+N-Topologie umfasst ein direkt mit einem gemeinsam genutzten Speicher verbundenes Knotenpaar und einen weiteren Knotensatz, der den Cluster-Interconnect für den Zugriff auf den gemeinsam genutzten Speicher einsetzt — sie sind nicht direkt miteinander verbunden.

Die nachstehende Abbildung stellt eine Paar+N-Topologie dar, in der zwei der vier Knoten (Knoten 3 und Knoten 4) den Cluster-Interconnect für den Speicherzugriff verwenden. Diese Konfiguration kann um zusätzliche Knoten erweitert werden, die keinen direkten Zugriff auf den gemeinsam genutzten Speicher haben.

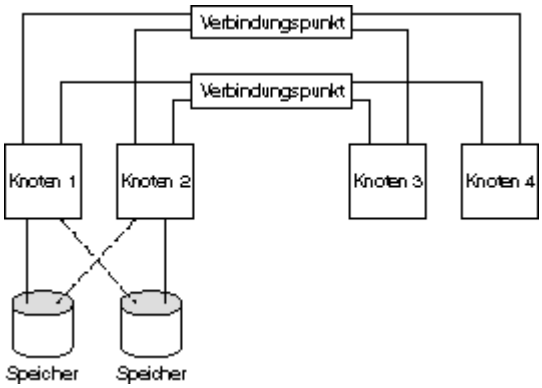


ABBILDUNG 2–4 Paar+N-Topologie

N+1- (Stern)-Topologie

Eine N+1-Topologie umfasst mehrere Primärknoten und einen Sekundärknoten. Sie müssen die Primärknoten und den Sekundärknoten nicht identisch konfigurieren. Die Primärknoten stellen aktiv Anwendungsdienste zur Verfügung. Der Sekundärknoten muss sich nicht im Leerlauf befinden, während er auf den Ausfall eines Primärknotens wartet.

Der Sekundärknoten ist der einzige Knoten in der Konfiguration, der real mit dem ganzen Multihostspeicher verbunden ist.

Wenn ein Fehler in einem Primärknoten auftritt, führt Sun Cluster ein Failover der Ressourcen auf den Sekundärknoten durch, auf dem die Ressourcen bleiben, bis sie (entweder automatisch oder manuell) wieder auf den Primärknoten umgeschaltet werden.

Der Sekundärknoten muss immer über eine ausreichende CPU-Überkapazität verfügen, um die Last übernehmen zu können, falls einer der Primärknoten ausfällt.

Die nachstehende Abbildung stellt eine N+1-Konfiguration dar.

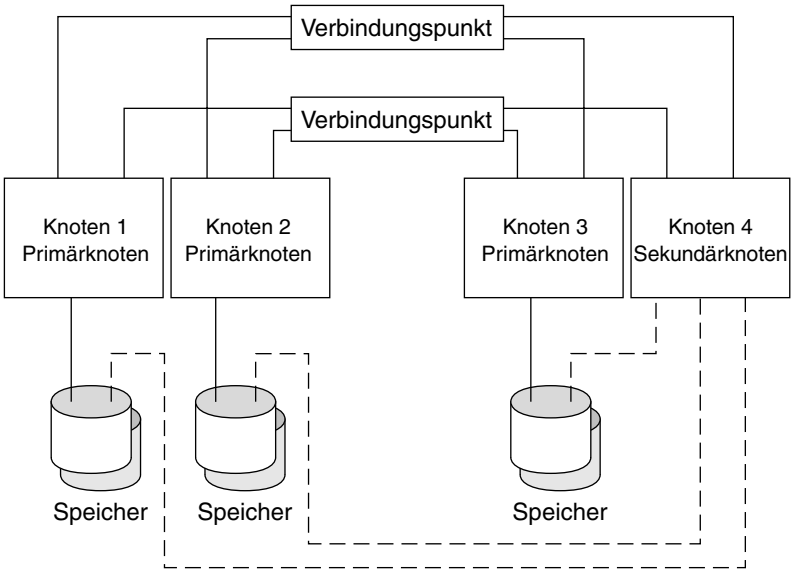


ABBILDUNG 2–5 N+1-Topologie

N*N-(Scalable)-Topologie

Mit einer N*N-Topologie können alle gemeinsam genutzten Speichergeräte in einem Cluster mit allen Knoten im Cluster verbunden werden. Mit dieser Topologie können hoch verfügbare Anwendungen ohne Abfall der Dienstqualität von einem Knoten auf einen anderen umgeleitet werden. Bei einem Failover kann der neue Knoten über einen lokalen Pfad auf das Speichergerät zugreifen, anstatt den privaten Interconnect zu verwenden.

Die nachstehende Abbildung stellt eine N*N-Konfiguration dar.

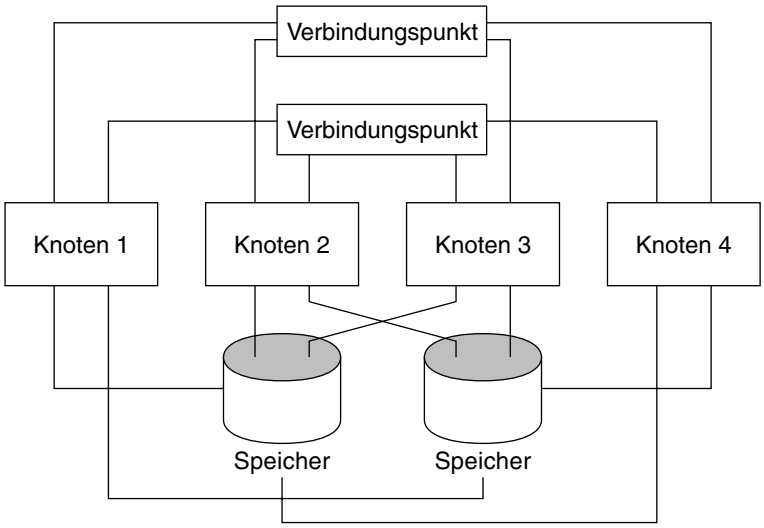


ABBILDUNG 2–6 N*N-Topologie

KAPITEL 3

Schlüsselkonzepte – Verwaltung und Anwendungsentwicklung

Dieses Kapitel beschreibt die Schlüsselkonzepte im Zusammenhang mit den Softwarekomponenten eines SunPlex-Systems. Zu den behandelten Themen gehören:

Cluster-Verwaltung und Anwendungsentwicklung

Diese Informationen richten sich in erster Linie an Systemverwalter und Anwendungsentwickler, die mit der SunPlex-API und dem SDK arbeiten. Cluster-Systemverwalter können diese Informationen als Hintergrund für das Installieren, Konfigurieren und Verwalten von Cluster-Software nutzen. Anwendungsentwickler können die Informationen nutzen, um die Cluster-Umgebung zu verstehen, in der sie arbeiten.

Die nachstehende Abbildung zeigt eine Übersicht auf höchster Ebene über die Zuordnung zwischen Cluster-Verwaltungskonzepten und Cluster-Architektur.

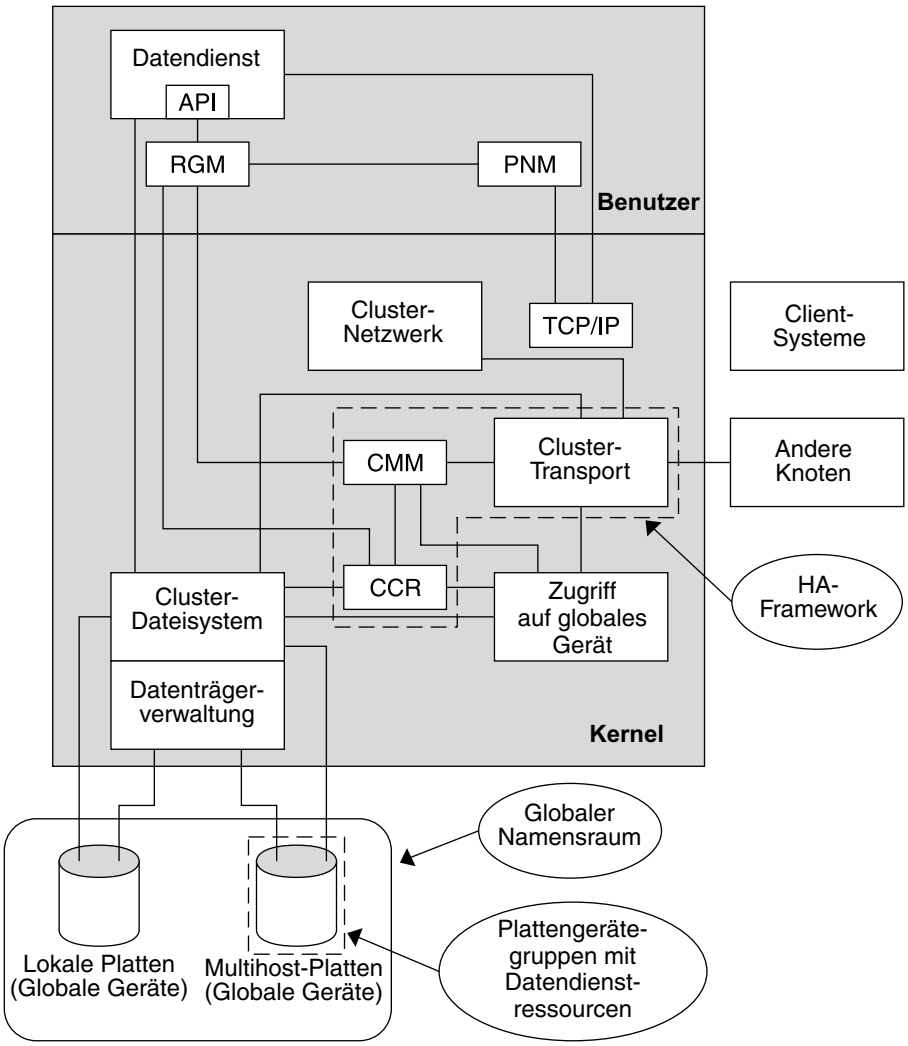


ABBILDUNG 3–1 Sun Cluster-Softwarearchitektur

Verwaltungsschnittstellen

Sie können auswählen, wie Sie das SunPlex-System von unterschiedlichen Benutzeroberflächen aus installieren, konfigurieren und verwalten möchten. Sie können Systemverwaltungsaufgaben entweder über die grafische Benutzeroberfläche

(Graphical User Interface, GUI) von SunPlex-Manager oder über die dokumentierte Befehlszeilenschnittstelle ausführen. Im obersten Bereich der Befehlszeilenschnittstelle befinden sich mehrere Dienstprogramme wie `scinstall` und `scsetup`, um bestimmte Installations- und Konfigurationsaufgaben zu vereinfachen. Zum SunPlex-System gehört auch ein Modul, das als Teil von Sun Management Center läuft und eine GUI für bestimmte Cluster-Aufgaben zur Verfügung stellt. Eine vollständige Beschreibung der Verwaltungsschnittstellen finden Sie im Einführungskapitel im *Sun Cluster 3.1 Handbuch Systemverwaltung*.

Cluster-Zeit

Die Zeit muss zwischen allen Knoten eines Clusters synchronisiert sein. Ob die Zeitsynchronisierung der Cluster-Knoten mit einer externen Zeitquelle erfolgt, hat für den Cluster-Betrieb keine Bedeutung. Das SunPlex-System verwendet das NTP-Protokoll (Network Time Protocol) zum Synchronisieren der Uhren zwischen den Knoten.

Im Allgemeinen verursacht eine Änderung der Systemuhr um eine Sekundenfraktion keine Probleme. Wenn Sie jedoch `date(1)`, `rdate(1M)` oder `xntpdate(1M)` (interaktiv oder innerhalb von `cron`-Skripts) auf einem aktiven Cluster ausführen, können Sie eine weit über einer Sekundenfraktion liegende Zeitänderung erzwingen, um die Systemuhr mit der Zeitquelle zu synchronisieren. Diese erzwungene Änderung kann zu Problemen bei den Zeitmarken für Dateiänderungen oder beim NTP-Dienst führen.

Wenn Sie die Solaris-Betriebsumgebung auf jedem Cluster-Knoten installieren, können Sie die Standardeinstellung für Zeit und Datum für den Knoten ändern. Im Allgemeinen können Sie die werkseitigen Standardeinstellungen akzeptieren.

Wenn Sie die Sun Cluster-Software mit `scinstall(1M)` installieren, wird in einem Prozessschritt auch das NTP für den Cluster konfiguriert. Die Sun Cluster-Software stellt eine Vorlagendatei zur Verfügung, `ntp.cluster` (siehe `/etc/inet/ntp.cluster` auf einem installierten Cluster-Knoten), die eine Peer-Beziehung zwischen allen Cluster-Knoten mit einem Knoten als “bevorzugtem” Knoten festlegt. Knoten werden durch ihre privaten Hostnamen identifiziert und die Zeitsynchronisierung erfolgt über den Cluster-Interconnect. Die Anweisungen zur Cluster-Konfiguration für NTP finden Sie im *Sun Cluster 3.1 Handbuch Softwareinstallation*.

Alternativ können Sie einen oder mehrere NTP-Server außerhalb des Clusters einrichten und die `ntp.conf`-Datei ändern, um diese Konfiguration wiederzugeben.

Im Normalbetrieb sollten Sie die Cluster-Zeit niemals anpassen müssen. Wenn die Zeit jedoch bei der Installation der Solaris-Betriebsumgebung falsch eingestellt wurde und Sie diese ändern möchten, finden Sie die Beschreibung der Vorgehensweise im *Sun Cluster 3.1 Handbuch Systemverwaltung*.

Hoch verfügbares Framework

Das SunPlex-System macht alle Komponenten auf dem “Pfad” zwischen Benutzern und Daten hoch verfügbar, einschließlich der Netzwerkschnittstellen, der Anwendungen selbst, des Dateisystems und der Multihostplatten. Im Allgemeinen ist eine Cluster-Komponente hoch verfügbar, wenn sie trotz eines einzelnen Fehlers (Software oder Hardware) im System in Betrieb bleibt.

Die nachstehende Tabelle zeigt die Arten von SunPlex-Komponentenfehlern (bei Hardware und Software) und die entsprechende Form der Wiederherstellung, die in das hoch verfügbare Framework integriert ist.

TABELLE 3–1 Ebenen der SunPlex-Fehlererkennung und Wiederherstellung

Fehlerhafte Cluster-Komponente	Softwarewiederherstellung	Hardwarewiederherstellung
Datendienst	HA-API, HA-Framework	Nicht verfügbar
Öffentlicher Netzwerkadapter	IP Network Multipathing	Mehrfache öffentliche Netzwerkadapterkarten
Cluster-Dateisystem	Primär- und Sekundärknotenreplikate	Multihostplatten
Gespiegelte Multihostplatte	Datenträgerverwaltung (Solaris Volume Manager und VERITAS Volume Manager)	RAID-5-Hardware (zum Beispiel Sun StorEdge™ A3x00)
Globales Gerät	Primär- und Sekundärknotenreplikate	Mehrere Pfade zum Gerät, Cluster-Transportverbindungspunkte
Privates Netzwerk	HA-Transportsoftware	Mehrere private, hardwareunabhängige Netzwerke
Knoten	CMM, Failfast-Treiber	Mehrere Knoten

Das Hochverfügbarkeits-Framework der Sun Cluster-Software erkennt einen Knotenfehler schnell und erstellt einen neuen, gleichwertigen Server für die Framework-Ressourcen auf einem anderen Knoten im Cluster. Zu keiner Zeit sind alle Framework-Ressourcen nicht verfügbar. Framework-Ressourcen, die von einem abgestürzten Knoten nicht betroffen sind, sind während der Wiederherstellung voll verfügbar. Darüber hinaus sind die Framework-Ressourcen des ausgefallenen Knotens wieder verfügbar, sobald sie wiederhergestellt sind. Eine wiederhergestellte Framework-Ressource muss nicht warten, bis die Wiederherstellung aller anderen Framework-Ressourcen abgeschlossen ist.

Die meisten hoch verfügbaren Framework-Ressourcen werden für die Anwendungen (Datendienste), die sie verwenden, transparent wiederhergestellt. Die Semantik des Framework-Ressourcenzugriffs bleibt bei einem Knotenfehler vollständig erhalten. Die Anwendungen nehmen einfach nicht wahr, dass der Framework-Ressourcenserver auf

einen anderen Knoten verschoben wurde. Das Versagen eines einzigen Knotens ist für die Programme auf den restlichen Knoten vollständig transparent, die so lange die mit diesem Knoten verbundenen Dateien, Geräte und Datenträger verwenden, wie ein alternativer Hardwarepfad zu den Platten eines anderen Knotens vorhanden ist. Ein Beispiel ist die Verwendung von Multihostplatten mit Ports für mehrere Knoten.

Cluster-Mitglieder-Monitor (CMM)

Der Cluster-Mitglieder-Monitor (CMM) ist ein verteilter Satz von Agenten mit einem Agenten pro Cluster-Mitglied. Die Agenten tauschen mit folgenden Zielsetzungen Meldungen über den Cluster-Interconnect aus:

- Erzwingen einer konsistenten Mitgliederansicht auf allen Knoten (Quorum),
- Durchführen einer synchronisierten Rekonfiguration infolge von Änderungen der Mitgliedschaft mithilfe von registrierten Rückmeldungen,
- Bearbeiten der Cluster-Partitionierung (Split Brain, Amnesie),
- Sicherstellen der vollen Konnektivität zwischen allen Cluster-Mitgliedern.

Im Unterschied zu früheren Sun Cluster-Softwareversionen wird der CMM vollständig im Kernel ausgeführt.

Cluster-Mitgliedschaft

Die Hauptfunktion des CMM ist das Festlegen einer Cluster-weiten Einigung über den Knotensatz, der zum jeweiligen Zeitpunkt am Cluster teilnimmt. Diese Einschränkung wird als *Cluster-Mitgliedschaft* bezeichnet.

Zur Feststellung der Cluster-Mitgliedschaft und letzten Endes zur Sicherung der Datenintegrität geht der CMM folgendermaßen vor:

- Er weist eine Änderung bei der Cluster-Mitgliedschaft aus, zum Beispiel ein Knoten, der dem Cluster beitritt oder diesen verlässt.
- Er stellt sicher, dass ein “fehlerhafter” Knoten den Cluster verlässt.
- Er stellt sicher, dass ein “fehlerhafter” Knoten außerhalb des Clusters bleibt, bis er repariert ist.
- Er verhindert, dass der Cluster sich selbst in Knoten-Teilsätze partitioniert.

Weitere Informationen darüber, wie sich der Cluster vor der Partitionierung in mehrere getrennte Cluster schützt, finden Sie unter „Quorum und Quorum-Geräte“ auf Seite 54.

Rekonfiguration des CMM

Alle Knoten müssen eine konsistente Einigung hinsichtlich der Cluster-Mitgliedschaft erzielen, um die Daten vor Beschädigung zu schützen. Der CMM koordiniert gegebenenfalls eine Cluster-Rekonfiguration der Cluster-Dienste (Anwendungen) infolge eines Fehlers.

Der CMM erhält Informationen über die Konnektivität mit anderen Knoten aus der Cluster-Transportschicht. Der CMM verwendet während einer Rekonfiguration den Cluster-Interconnect zum Austauschen von Statusinformationen.

Nach dem Erkennen einer Änderung bei der Cluster-Mitgliedschaft führt der CMM eine synchronisierte Konfiguration des Clusters aus, bei der die Cluster-Ressourcen ggf. auf Grundlage der neuen Mitgliedschaft im Cluster neu verteilt werden.

Failfast-Mechanismus

Wenn der CMM ein kritisches Knoten-Problem erkennt, fordert er das Cluster-Framework auf, den Knoten zwangsweise herunterzufahren (Panik) und aus der Cluster-Mitgliedschaft zu entfernen. Der Mechanismus für diesen Vorgang wird als *Failfast* bezeichnet. Ein Failfast bewirkt das Herunterfahren eines Knotens auf zwei Arten.

- Wenn ein Knoten den Cluster verlässt und dann versucht, einen neuen Cluster ohne Quorum zu starten, wird er vom Zugriff auf die gemeinsam genutzten Platten "geschützt". Weitere Details zur Verwendung von Failfast finden Sie unter „Fehlerschutz“ auf Seite 57.
- Wenn einer oder mehr Cluster-spezifische Dämonen ausfallen (`clexecd`, `rpc.pmfd`, `rgmd` oder `rpc.ed`), wird der Fehler vom CMM erkannt und der Knoten gerät in Panik. Wenn der Ausfall eines Cluster-Dämons den Knoten in Panik versetzt, wird in der Konsole für diesen Knoten eine Meldung angezeigt, die in etwa folgendermaßen aussieht.

```
panic[cpu0]/thread=40e60: Failfast: Aborting because "pmfd" dies 35 seconds ago.
409b8 cl_runtime:__0FZsc_syslog_msg_log_no_argsPviTCPCcTB+48 (70f900, 30, 70df54, 407acc, 0)
%l0-7: 1006c80 000000a 000000a 10093bc 406d3c80 7110340 0000000 4001 fbf0
```

Anschließend kann der Knoten neu booten und versuchen, wieder dem Cluster beizutreten, oder kann an der OpenBoot™ PROM (OBP)-Eingabeaufforderung bleiben. Die durchgeführte Aktion wird von der Einstellung des `auto-boot?`-Parameters im OBP bestimmt.

Cluster-Konfigurations-Repository (CCR)

Das Cluster-Konfigurations-Repository (CCR) ist eine private, Cluster-weite Datenbank zur Speicherung von Informationen über Konfiguration und Zustand des Clusters. Das CCR ist eine verteilte Datenbank. Jeder Knoten pflegt eine vollständige

Kopie der Datenbank. Das CCR stellt sicher, dass alle Knoten ein konsistentes Bild der Cluster-"Welt" haben. Um eine Beschädigung der Daten zu vermeiden, muss jeder Knoten den aktuellen Zustand der Cluster-Ressourcen kennen.

Das CCR verwendet für Aktualisierungen einen Zwei-Phasen-Commit-Algorithmus: Eine Aktualisierung muss auf allen Cluster-Mitgliedern erfolgreich abgeschlossen werden, sonst wird die Aktualisierung zurückgenommen. Das CCR verwendet den Cluster-Interconnect für die Umsetzung der verteilten Aktualisierungen.

Achtung – Das CCR besteht zwar aus Textdateien, aber Sie dürfen die CCR-Dateien unter keinen Umständen manuell bearbeiten. Jede Datei enthält einen Prüfsummeneintrag, um die Konsistenz zwischen den Knoten zu sichern. Eine manuell Aktualisierung der CCR-Dateien kann dazu führen, dass ein Knoten oder der ganze Cluster nicht mehr funktioniert.

Das CCR stellt mithilfe des CMM sicher, dass ein Cluster nur mit einem festgelegten Quorum läuft. Das CCR ist dafür zuständig, die Datenkonsistenz im ganzen Cluster zu überprüfen, eine ggf. erforderliche Wiederherstellung durchzuführen und Aktualisierungen für die Daten bereitzustellen.

Globale Geräte

Das SunPlex-System verwendet *globale Geräte*, um einen Cluster-weiten, hoch verfügbaren Zugriff auf alle Geräte innerhalb eines Clusters von jedem Knoten aus zur Verfügung zu stellen, und zwar unabhängig davon, wo das Gerät real angeschlossen ist. Wenn ein Knoten ausfällt, während er Zugriff auf ein globales Gerät bietet, erkennt die Sun Cluster-Software im Allgemeinen automatisch einen anderen Pfad zum Gerät und leitet den Zugriff auf diesen Pfad um. Zu den globalen Geräten bei SunPlex gehören Platten, CD-ROM und Bänder. Dabei sind die Platten jedoch die einzigen globalen Multiport-Geräte, die unterstützt werden. Das bedeutet, dass CD-ROM- und Bandgeräte derzeit keine hoch verfügbaren Geräte sind. Die lokalen Platten auf jedem Server sind ebenfalls keine Multiport-Geräte und deswegen nicht hoch verfügbar.

Der Cluster weist jeder Platte, jedem CD-ROM-Laufwerk und jedem Bandgerät im Cluster automatisch eine einmalige ID zu. Diese Zuweisung ermöglicht einen konsistenten Zugriff auf jedes Gerät von jedem Cluster-Knoten aus. Der Namensraum globaler Geräte ist im Verzeichnis `/dev/global` enthalten. Weitere Informationen finden Sie unter „Globaler Namensraum“ auf Seite 46.

Globale Multiport-Geräte stellen mehrere Pfade zu einem Gerät zur Verfügung. Da Multihostplatten zu einer Plattengerätegruppe gehören, die von mehreren Knoten gehostet wird, werden sie hoch verfügbar gemacht.

Geräte-ID (DID)

Die Sun Cluster-Software verwaltet globale Geräte über ein Gebilde, das als DID-Pseudotreiber (Device ID, DID) bezeichnet wird. Dieser Treiber wird verwendet, um automatisch jedem Gerät im Cluster einschließlich Multihostplatten, Bandlaufwerken und CD-ROM-Laufwerken eine einmalige ID zuzuweisen.

Der DID-Pseudotreiber ist fester Bestandteil der Zugriffsfunktion auf globale Geräte des Clusters. Der DID-Treiber testet alle Knoten des Clusters, erstellt eine Liste von einmaligen Plattengeräten und weist jeder Platte eine einmalige Geräteklassen- und Gerätenummer zu, die auf allen Knoten des Clusters konsistent ist. Der Zugriff auf globale Geräte erfolgt mithilfe der einmaligen, vom DID-Treiber zugewiesenen Geräte-ID anstelle der herkömmlichen Solaris-Geräte-IDs wie `c0t0d0` für eine Platte.

Dieser Ansatz stellt sicher, dass jede auf Platten zugreifende Anwendung (wie Datenträger-Manager oder Anwendungen, die im raw-Modus betriebene Geräte verwenden) im ganzen Cluster einen konsistenten Pfad verwenden. Diese Konsistenz ist besonders bei Multihostplatten wichtig, weil die lokale Geräteklassen- und Gerätenummern für jedes Gerät von Knoten zu Knoten unterschiedlich sein kann und sich damit auch die Solaris-Konventionen zur Gerätebenennung ändern. Knoten1 kann eine Multihostplatte zum Beispiel als `c1t2d0` führen, während Knoten2 dieselbe Platte vollkommen anders, nämlich als `c3t2d0` , führt. Der DID-Treiber weist einen globalen Namen zu, wie d10, den der Knoten statt dessen verwendet, und gibt jedem Knoten eine konsistente Zuordnung zur Multihostplatte.

Sie aktualisieren und verwalten die Geräte-IDs mithilfe von `scdidadm(1M)` und `scgdevs(1M)`. Weitere Informationen finden Sie in der jeweiligen Online-Dokumentation.

Plattengerätegruppen

Im SunPlex-System müssen alle Multihostplatten von der Sun Cluster-Software gesteuert werden. Sie erstellen zunächst die Plattengruppen des Datenträger-Managers — entweder Solaris Volume Manager-Plattensätze oder VERITAS Volume Manager-Plattengruppen — auf den Multihostplatten. Dann registrieren Sie die Datenträger-Manager-Plattengruppen als *Plattengerätegruppen*. Eine Plattengerätegruppe ist ein globaler Gerätetyp. Zudem erstellt die Sun Cluster-Software automatisch eine im raw-Modus betriebene Gerätegruppe für jedes Platten- und Bandgerät im Cluster. Diese Cluster-Gerätegruppen bleiben jedoch im Offline-Zustand, bis Sie darauf als globale Geräte zugreifen.

Die Registrierung liefert dem SunPlex-System Informationen zu den Pfaden zwischen Knoten und Datenträger-Manager-Plattengruppen. An diesem Punkt werden Datenträger-Manager-Plattengruppen innerhalb des Clusters global zugänglich. Wenn mehrere Knoten in eine Plattengerätegruppe schreiben können (die Gruppe unterstützen können), werden die in dieser Plattengerätegruppe gespeicherten Daten hoch verfügbar. Die hoch verfügbare Plattengerätegruppe kann zum Hosten von Cluster-Dateisystemen verwendet werden.

Hinweis – Plattengerätegruppen sind von Ressourcengruppen unabhängig. Ein Knoten kann eine Ressourcengruppe (die eine Gruppe von Datendienstprozessen darstellt) unterstützen, während ein anderer die Plattengruppe(n) unterstützen kann, auf welche die Datendienste zugreifen. Das optimale Vorgehen besteht aber darin, die Plattengerätegruppe zur Speicherung bestimmter Anwendungsdaten und die Ressourcengruppe mit den Anwendungsressourcen (Anwendungsdämon) auf demselben Knoten zu speichern. Weitere Informationen zur Zuordnung von Plattengerätegruppen und Ressourcengruppen finden Sie im Überblickskapitel im *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Mit einer Plattengerätegruppe wird die Datenträger-Manager-Plattengruppe “global”, weil sie Multipathing für die dazugehörigen Platten unterstützt. Jeder mit den Multihostplatten real verbundener Cluster-Knoten stellt einen Pfad zur Plattengerätegruppe zur Verfügung.

Plattengerätegruppen-Failover

Auf alle Plattengerätegruppen in einem Gehäuse kann über einen alternativen Pfad zugegriffen werden, wenn der derzeitige Knoten-Master der Gerätegruppe ausfällt, weil ein Plattengehäuse an mehrere Knoten angeschlossen ist. Der Ausfall des Knotens, der als Gerätegruppen-Master fungiert, beeinträchtigt den Zugriff auf die Gerätegruppe nur während der Zeitspanne, die zur Ausführung der Wiederherstellungs- und Konsistenzprüfung erforderlich ist. Während dieser Zeit werden alle Anforderungen (für die Anwendung transparent) gesperrt, bis das System die Gerätegruppe verfügbar macht.

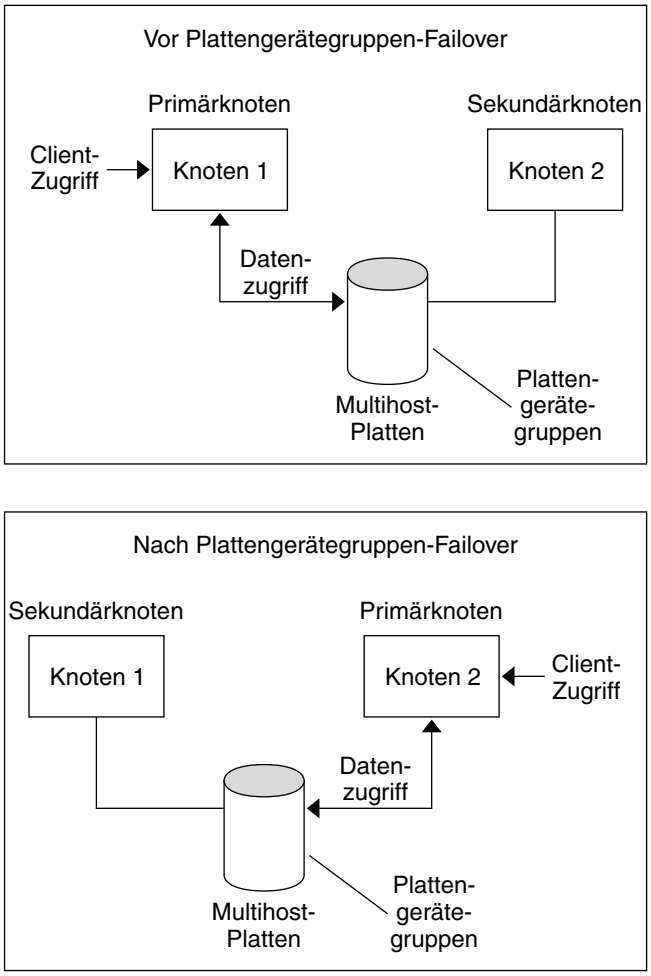


ABBILDUNG 3–2 Plattengerätegruppen-Failover

Multiport-Plattengerätegruppen

In diesem Abschnitt werden Plattengruppeneigenschaften beschrieben, mit denen Sie Leistung und Verfügbarkeit in einer Multiport-Plattenkonfiguration ausgleichen können. Die Sun Cluster-Software stellt Ihnen zwei Eigenschaften zur Verfügung, mit denen Sie eine Multiport-Plattenkonfiguration konfigurieren können: `preferenced` und `numsecondaries` . Mit der `preferenced`-Eigenschaft können Sie die Reihenfolge steuern, in der die Knoten versuchen, die Steuerung im Fall eines Failover zu übernehmen. Mit der `numsecondaries`-Eigenschaft legen Sie eine gewünschte Anzahl von Sekundärknoten für eine Gerätegruppe fest.

Ein hoch verfügbarer Dienst gilt als ausgefallen, wenn der Primärknoten ausfällt und kein Sekundärknoten die Rolle des Primärknotens übernehmen kann. Wenn ein Dienst-Failover erfolgt und die `preferenced`-Eigenschaft auf `true` eingestellt ist, folgen die Knoten bei der Auswahl eines Sekundärknotens der Reihenfolge in der Knotenliste. Die mit der preferenced-Eigenschaft eingerichtete Knotenliste definiert die Reihenfolge, in der die Knoten versuchen, die Steuerung als Primärknoten oder den Übergang von Spare-Knoten zu Sekundärknoten zu übernehmen. Sie können den Vorrang eines Gerätedienstes mit dem Dienstprogramm `scsetup`(1M) dynamisch ändern. Der Vorrang der abhängigen Dienste, zum Beispiel eines globalen Dateisystems, entspricht dem des Gerätedienstes.

Im Normalbetrieb führt der Primärknoten Checkpoint-Vorgänge für Sekundärknoten durch. In einer Multiport-Plattenkonfiguration führt das Durchführen des Checkpoint-Vorgangs an jedem Sekundärknoten zu einer Leistungsverminderung des Clusters und zu Speicherüberlastung. Die Spare-Knoten-Unterstützung wurde implementiert, um die Checkpoint-bedingte Leistungsminderung und Speicherüberlastung zu minimieren. Standardmäßig hat Ihre Plattengerätegruppe einen Primär- und einen Sekundärknoten. Die restlichen verfügbaren Anbieterknoten werden im Spare-Zustand online gebracht. Bei einem Failover wird der Sekundärknoten zum Primärknoten, und der Knoten mit der höchsten Priorität in der Knotenliste wird zum Sekundärknoten.

Die gewünschte Anzahl von Sekundärknoten kann auf jede beliebige Ganzzahl zwischen Eins und der Anzahl der betriebsbereiten Anbieterknoten in der Gerätegruppe, die keine Primärknoten sind, gesetzt werden.

Hinweis – Wenn Sie Solaris Volume Manager verwenden, müssen Sie zuerst die Plattengerätegruppe erstellen, bevor Sie die Eigenschaft `numsecondaries` auf einen anderen Wert als den Standardwert einstellen können.

Die standardmäßig eingestellte gewünschte Anzahl von Sekundärknoten für Gerätedienste ist Eins. Die tatsächliche Anzahl der vom Replikat-Framework verwalteten Sekundäranbieter entspricht der gewünschten Anzahl, es sei denn, die Anzahl der betriebsbereiten Anbieter, die keine Primärknoten sind, ist kleiner als die gewünschte Anzahl. Sie möchten ggf. die `numsecondaries`-Eigenschaft ändern und die Knotenliste querprüfen, wenn Sie der Konfiguration Knoten hinzufügen oder daraus entfernen. Die Verwaltung der Knotenliste und der gewünschten Anzahl von Sekundärknoten verhindert Konflikte zwischen der konfigurierten und der tatsächlichen, vom Framework erlaubten Anzahl von Sekundärknoten. Mit dem `scconf`(1M)-Befehl für VxVM-Plattengerätegruppen oder dem `metaset`(1M)-Befehl für Solaris Volume Manager-Gerätegruppen zusammen mit der Einstellung der Eigenschaften `preferenced` und `numsecondaries` verwalten Sie das Hinzufügen oder Entfernen von Knoten aus Ihrer Konfiguration. Informationen zum Vorgehen bei der Änderung der Eigenschaften von Plattengerätegruppen finden Sie unter " Administering Global Devices and Cluster File Systems" in *Sun Cluster 3.1 10/03 Handbuch Systemverwaltung*.

Globaler Namensraum

Der Sun Cluster-Softwaremechanismus zur Aktivierung globaler Geräte ist der *globale Namensraum*. Der globale Namensraum beinhaltet die `/dev/global/`-Hierarchie ebenso wie die Datenträger-Manager-Namensräume. Der globale Namensraum gibt sowohl Multihostplatten als auch lokale Platten wieder (und jedes andere Cluster-Gerät wie CD-Rom-Laufwerke und Bänder) und stellt mehrere Failover-Pfade für die Multihostplatten zur Verfügung. Jeder real an die Multihostplatten angeschlossene Knoten stellt für alle Knoten im Cluster einen Pfad zum Speicher zur Verfügung.

In der Regel befinden sich die Namensräume der Datenträger-Manager in den Verzeichnissen `/dev/md/`*Plattensatz*`/dsk` (und `rdsk`) für den Solaris Volume Manager und in den Verzeichnissen `/dev/vx/dsk/` *Plattengruppe* und `/dev/vx/rdsk/` *Plattengruppe* für den VxVM. Diese Namensräume bestehen aus Verzeichnissen für jeden Solaris Volume Manager-Plattensatz bzw. jede VxVM-Plattengruppe auf dem ganzen Cluster. Jedes dieser Verzeichnisse enthält einen Geräteknoten für jedes Metagerät oder jeden Datenträger in diesem Plattensatz bzw. in dieser Plattengruppe.

Im SunPlex-System wird jeder Geräteknoten im lokalen Datenträger-Manager-Namensraum durch eine symbolische Verknüpfung mit einem Geräteknoten im Dateisystem `/global/.devices/node@` *Knoten-ID* ersetzt, wobei *Knoten-ID* als Ganzzahl für den Knoten im Cluster steht. Die Sun Cluster-Software zeigt die Datenträger-Manager-Geräte auch weiterhin als symbolische Verknüpfungen an ihren Standard-Speicherorten an. Sowohl der globale Namensraum als auch der Datenträger-Manager-Namensraum sind auf jedem Cluster-Knoten verfügbar.

Zu den Vorteilen des globalen Namensraumes gehört Folgendes:

- Jeder Knoten bleibt ziemlich unabhängig, und es gibt nur geringfügige Änderungen im Geräteverwaltungsmodell.
- Geräte können selektiv global gemacht werden.
- Verknüpfungsgeneratoren von Drittherstellern arbeiten weiterhin.
- Nach Vergabe eines lokalen Gerätenamens erfolgt eine einfache Zuordnung, um den entsprechenden globalen Namen zu erhalten.

Beispiel für lokale und globale Namensräume

Die nachstehende Tabelle zeigt die Zuordnungen zwischen lokalen und globalen Namensräumen für eine Multihostplatte, `c0t0d0s0`.

TABELLE 3–2 Zuordnungen von lokalen und globalen Namensräumen

Komponente/Pfad	Lokaler Knoten-Namensraum	Globaler Namensraum
Logischer Solaris-Name	`/dev/dsk/c0t0d0s0`	`/global/.devices/node@`*Knoten-ID* `/dev/dsk/c0t0d0s0`
DID-Name	`/dev/did/dsk/d0s0`	`/global/.devices/node@`*Knoten-ID* `/dev/did/dsk/d0s0`
Solaris Volume Manager	`/dev/md/`*Plattensatz*`/dsk/d0`	`/global/.devices/nodes@` *Knoten-ID*`/dev/md/`*Plattensatz*`/dsk/d0`
VERITAS Volume Manager	`/dev/vx/dsk/`*Plattengruppe*`/v0`	`/global/.devices/node@` *Knoten-ID*`/dev/vx/dsk/`*Plattengruppe*`/v0`

Der globale Namensraum wird automatisch bei der Installation erzeugt und bei jedem Neubooten der Rekonfiguration aktualisiert. Sie können den globalen Namensraum auch durch Ausführen des `scgdevs(1M)`-Befehls generieren.

Cluster-Dateisysteme

Ein Cluster-Dateisystem ist ein Proxy zwischen dem Kernel auf einem Knoten und dem zugrunde liegenden Dateisystem und dem Datenträger-Manager auf einem anderen Knoten mit einem realen Anschluss an die Platte(n).

Cluster-Dateisysteme hängen von globalen Geräten (Platten, Bänder, CD-ROMs) mit realen Verbindungen mit mindestens einem Knoten ab. Sie können von jedem Knoten im Cluster über denselben Dateinamen (zum Beispiel `/dev/global/`) auf die globalen Geräte zugreifen, unabhängig davon, ob dieser Knoten tatsächlich mit dem Speichergerät verbunden ist. Sie können ein globales Gerät genauso verwenden wie ein normales Gerät, das heißt, Sie können mit den Befehlen `newfs` und/oder `mkfs` Dateisysteme darauf erstellen.

Sie können ein Dateisystem mit `mount -g` global oder mit `mount` lokal auf einem globalen Gerät einhängen.

Programme können auf die Dateien in einem Cluster-Dateisystem von jedem Knoten im Cluster aus und mit demselben Dateinamen zugreifen (zum Beispiel `/global/foo`).

Ein Cluster-Dateisystem wird auf allen Cluster-Mitgliedern eingehängt. Sie können ein Cluster-Dateisystem nicht auf einer Untermenge von Cluster-Mitgliedern einhängen.

Ein Cluster-Dateisystem ist kein anderer Typ von Dateisystem. Das bedeutet, der Client sieht das zugrunde liegende Dateisystem (zum Beispiel UFS).

Verwenden von Cluster-Dateisystemen

Im SunPlex-System sind alle Multihostplatten in Plattengerätegruppen eingebunden. Hierbei kann es sich um Solaris Volume Manager-Plattensätze, VxVM-Plattengruppen oder einzelne Platten handeln, die nicht von einem softwarebasierten Datenträger-Manager gesteuert werden.

Damit ein Cluster-Dateisystem hoch verfügbar ist, muss der zugrunde liegende Plattenspeicher mit mehreren Knoten verbunden sein. Deswegen ist ein lokales Dateisystem (ein auf der lokalen Platte eines Knotens gespeichertes Dateisystem), das zu einem Cluster-Dateisystem gemacht wird, nicht hoch verfügbar.

Wie bei normalen Dateisystemen können Sie auch Cluster-Dateisysteme auf zwei Arten einhängen:

- **Manuell** — Verwenden Sie den `mount`-Befehl mit den Einhängeoptionen `-g` oder `-o global`, um das Cluster-Dateisystem von der Befehlszeile aus einzuhängen. Beispiel:

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

- **Automatisch** — Erstellen Sie einen Eintrag in der Datei `/etc/vfstab` mit einer `global`-Einhängeoption, um das Cluster-Dateisystem beim Booten einzuhängen. Dann können Sie im `/global`-Verzeichnis auf allen Knoten einen Einhängepunkt erstellen. Das `/global`-Verzeichnis ist ein empfohlener Speicherplatz, keine Anforderung. Hier eine Beispielzeile für ein Cluster-Dateisystem aus einer `/etc/vfstab`-Datei:

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

Hinweis – Solange die Sun Cluster-Software kein Benennungsverfahren für Cluster-Dateisysteme vorschreibt, können Sie die Verwaltung durch das Erstellen eines Einhängepunktes für alle Cluster-Dateisysteme unter demselben Verzeichnis vereinfachen, wie `/global/`*Plattengerätegruppe*. Weitere Informationen finden Sie im *Sun Cluster 3.1 Handbuch Softwareinstallation* und im *Sun Cluster 3.1 Handbuch Systemverwaltung*.

Merkmale von Cluster-Dateisystemen

Das Cluster-Dateisystem hat folgende Merkmale:

- Die Speicherorte für den Dateizugriff sind transparent. Ein Prozess kann eine Datei an einem beliebigen Speicherort im System öffnen, und die Prozesse können die Datei auf allen Knoten mithilfe desselben Pfadnamens suchen.

Hinweis – Wenn das Cluster-Dateisystem Dateien liest, wird die Zugriffszeit für diese Dateien nicht aktualisiert.

- Mit Kohärenzprotokollen wird die UNIX-Dateizugriffssemantik auch dann bewahrt, wenn mehrere Knoten gleichzeitig auf die Datei zugreifen.
- Umfassendes Caching wird zusammen mit Bulk-E/A-Bewegungen ohne Zwischenspeicherung (Zero-Copy) eingesetzt, um Dateidaten effizient zu verschieben.
- Das Cluster-Dateisystem stellt mit den `fcntl(2)`-Schnittstellen eine hoch verfügbare kooperative Dateisperrfunktion zur Verfügung. Anwendungen, die auf mehreren Knoten ausgeführt werden, können den Datenzugriff unter Verwendung der kooperativen Dateisperrung für eine Cluster-Dateisystem-Datei synchronisieren. Dateisperren werden sofort von allen den Cluster verlassenden Knoten und von allen gesperrten und fehlgeschlagenen Anwendungen aufgehoben.
- Kontinuierlicher Datenzugriff ist auch bei Ausfällen gesichert. Anwendungen sind von den Ausfällen nicht betroffen, solange noch ein Pfad zu den Platten funktionsfähig ist. Diese Garantie gilt für den Plattenzugriff im raw-Modus und für alle Dateisystemvorgänge.
- Cluster-Dateisysteme sind vom zugrunde liegenden Dateisystem und von der Datenträgerverwaltungs-Software unabhängig. Cluster-Dateisysteme machen jedes unterstützte, auf den Platten vorhandene Dateisystem global.

`HAStoragePlus`-Ressourcentyp

Der `HAStoragePlus`-Ressourcentyp ist darauf ausgelegt, nicht globale Dateisystemkonfigurationen wie UFS und VxFS hoch verfügbar zu machen. Mit `HAStoragePlus` integrieren Sie das lokale Dateisystem in die Sun Cluster-Umgebung und machen das Dateisystem hoch verfügbar. `HAStoragePlus` stellt zusätzliche Dateisystemfunktionen wie Prüfen, Einhängen und erzwungenes Aushängen zur Verfügung, mit denen Sun Cluster bei lokalen Dateisystemen ein Failover durchführen kann. Zum Ausführen eines Failover-Vorgangs muss sich das lokale Dateisystem auf globalen Plattengruppen mit aktivierten Affinitäts-Switchover befinden.

In den jeweiligen Datendienstkapiteln im *Data Services Installation and Configuration Guide* oder unter “Enabling Highly Available Local File Systems” in Kapitel 14, “Administering Data Services Resources”, finden Sie Informationen zur Verwendung des `HAStoragePlus`-Ressourcentyps.

`HAStoragePlus` kann auch zum Synchronisieren beim Start von Ressourcen und Plattengerätegruppen, von denen die Ressourcen abhängen, eingesetzt werden. Weitere Informationen finden Sie unter „Ressourcen, Ressourcengruppen und Ressourcentypen“ auf Seite 72.

Syncdir-Einhängeoption

Die `syncdir`-Einhängeoption kann für auf UFS basierende Cluster-Dateisysteme verwendet werden. Sie erzielen jedoch eine wesentliche bessere Leistung, wenn Sie `syncdir` nicht angeben. Bei Angabe von `syncdir` sind die Einträge garantiert POSIX-konform. Wird die Option nicht angegeben, verhält sich das System wie ein NFS-Dateisystem. In einigen Fällen würden Sie ohne `syncdir` fehlende Speicherkapazität zum Beispiel erst beim Schließen der Datei feststellen. Mit `syncdir` (und POSIX-Verhalten) wäre das Fehlen von Speicherplatz während des Schreibvorgangs erkannt worden. Es treten selten Probleme auf, wenn Sie `syncdir` nicht angeben. Wir empfehlen Ihnen deswegen, auf die Angabe zu verzichten und den Leistungsgewinn zu nutzen.

VxFS hat keine Einhängeoption, die der `syncdir` -Einhängeoption für UFS entspricht. Das VxFS-Verhalten entspricht dem UFS-Verhalten, wenn die `syncdir`-Einhängeoption nicht angegeben wurde.

Häufig gestellte Fragen zu globalen Geräten und Cluster-Dateisystemen finden Sie unter „Häufige Fragen zu Dateisystemen“ auf Seite 92.

Plattenpfadüberwachung

Die derzeitige Version der Sun Cluster-Software unterstützt die Plattenpfadüberwachung (DPM). Dieser Abschnitt liefert konzeptionelle Informationen zu DPM, zum DPM-Dämon und zu den Verwaltungstools für die Plattenpfadüberwachung. Verfahrenstechnische Information zur Überwachung, zum Aufheben der Überwachung und zum Prüfen des Plattenpfadstatus finden Sie im *Sun Cluster 3.1 10/03 Handbuch Systemverwaltung*.

Hinweis – DPM wird auf Knoten mit Versionen vor Sun Cluster 3.1 5/03 software nicht unterstützt. Verwenden Sie während einer Aufrüstung keine DPM-Befehle. Nach der Aufrüstung müssen alle Knoten online sein, um die DPM-Befehle verwenden zu können.

Überblick

DPM verbessert die gesamte Zuverlässigkeit der Failover- und Switchover-Vorgänge durch die Überwachung der Verfügbarkeit des Plattenpfades für den Sekundärknoten. Mit dem `scdpm`-Befehl überprüfen Sie die Verfügbarkeit des von einer Ressource verwendeten Plattenpfades, bevor die Ressource umgeschaltet wird. Mit den vom `scdpm`-Befehl zur Verfügung gestellten Optionen können Sie die Plattenpfade zu einem einzelnen oder zu allen Knoten im Cluster überwachen. Weitere Informationen zu Befehlszeilenoptionen finden Sie unter `scdpm`(1M).

Die DPM-Komponenten werden aus dem SUNWscu-Paket installiert. Das SUNWscu-Paket wird im Rahmen der Sun Cluster-Standardinstallation installiert. Details zur Installationsoberfläche finden Sie in der Online-Dokumentation scinstall(1M). In der nachstehenden Tabelle wird der Standardspeicherort für die Installation der DPM-Komponenten beschrieben.

Speicherort	Komponente
Dämon	/usr/cluster/lib/sc/scdpmd
Befehlszeilenschnittstelle	/usr/cluster/bin/scdpm
Gemeinsam genutzte Bibliotheken	/user/cluster/lib/libscdpm.so
Dämon-Statusdatei (zur Laufzeit erstellt)	/var/run/cluster/scdpm.status

Ein Multithread-DPM-Dämon wird auf jedem Knoten ausgeführt. Der DPM-Dämon (scdpmd) wird durch ein rc.d-Skript gestartet, wenn ein Knoten bootet. Wenn ein Problem auftritt, wird der Dämon von pmfd verwaltet und startet automatisch neu. In der nachstehenden Liste wird die Arbeitsweise von scdpmd beim ersten Start beschrieben.

Hinweis – Beim Start wird der Status für jeden Plattenpfad auf UNKNOWN initialisiert.

---

1. Der DPM-Dämon sammelt Plattenpfad- und Knotennameninformationen aus der vorherigen Statusdatei oder aus der CCR-Datenbank. Weitere Informationen zum CCR finden Sie unter „Cluster-Konfigurations-Repository (CCR)“ auf Seite 40. Nach dem Start eines DPM-Dämons können Sie den Dämon zwingen, die Liste der überwachten Platten aus einer angegebenen Datei zu lesen.
2. Der DPM-Dämon initialisiert die Kommunikationsschnittstelle, um Anforderungen von Komponenten zu beantworten, die außerhalb des Dämons liegen, wie die Befehlszeilenschnittstelle.
3. Der DPM-Dämon pingt jeden Plattenpfad der überwachten Liste alle 10 Minuten mit dem Befehl scsi_inquiry an. Die Einträge werden gesperrt, um die Kommunikationsschnittstelle daran zu hindern, auf einen Eintrag zuzugreifen, der gerade geändert wird.
4. Der DPM-Dämon benachrichtigt das Sun Cluster-Ereignis-Framework und protokolliert den neuen Pfadstatus über den UNIX syslogd(1M)-Mechanismus.

Hinweis – Alle mit dem Dämon zusammenhängenden Fehler werden von pmfd(1M) aufgezeichnet. Alle API-Funktionen geben 0 für erfolgreich und -1 für einen Fehler zurück.

Der DPM-Dämon überwacht die Verfügbarkeit des logischen Pfades, der über Multipath-Treiber wie MPxIO, HDLM und PowerPath sichtbar ist. Die jeweiligen realen Pfade, die von diesen Treibern verwaltet werden, werden nicht überwacht, weil der Multipath-Treiber einzelne Ausfälle vor dem DPM-Dämon verbirgt.

Überwachen von Plattenpfaden

In diesem Abschnitt werden zwei Methoden zur Überwachung von Plattenpfaden in Ihrem Cluster beschrieben. Die erste Methode wird mit dem `scdpm`-Befehl zur Verfügung gestellt. Mit diesem Befehl können Sie die Plattenpfade im Cluster überwachen, die Überwachung aufheben oder deren Status anzeigen. Dieser Befehl dient auch dem Drucken der Liste von ausgefallenen Platten und der Überwachung der Plattenpfade aus einer Datei.

Die zweite Methode zur Überwachung von Plattenpfaden im Cluster wird von der grafischen Benutzeroberfläche (Grafical User Interface, GUI) von SunPlex-Manager zur Verfügung gestellt. SunPlex-Manager gibt Ihnen einen topologischen Überblick über die überwachten Plattenpfade im Cluster. Die Ansicht wird alle 10 Minuten aktualisiert, um Informationen zur Anzahl der fehlgeschlagenen Pings zu liefern. Zur Verwaltung der Plattenpfade verwenden Sie die Informationen der GUI von SunPlex-Manager zusammen mit dem `scdpm`(1M)-Befehl. Informationen zu SunPlex-Manager finden Sie unter " Administering Sun Cluster With the Graphical User Interfaces" in *Sun Cluster 3.1 10/03 Handbuch Systemverwaltung*.

Verwenden des `scdpm`-Befehls zur Plattenpfadüberwachung

Der `scdpm`(1M)-Befehl stellt DMP-Verwaltungsbefehle zur Verfügung, mit denen Sie folgende Aufgaben ausführen können:

- Überwachen eines neuen Plattenpfades,
- Aufheben der Überwachung eines Plattenpfades,
- Neues Lesen der Konfigurationsdaten aus der CCR-Datenbank,
- Lesen der Platten, die überwacht werden sollen oder deren Überwachung aufgehoben werden soll, von einer angegebenen Datei aus,
- Berichten des Status eines Plattenpfades oder aller Plattenpfade im Cluster,
- Drucken aller Plattenpfade, auf die von einem Knoten aus zugegriffen werden kann.

Geben Sie den `scdpm`(1M)-Befehl mit dem Plattenpfad-Argument von einem beliebigen aktiven Knoten aus, um DPM-Verwaltungsaufgaben auf dem Cluster auszuführen. Das Plattenpfad-Argument besteht immer aus einem Knotennamen und einem Plattennamen. Der Knotenname ist nicht erforderlich und wird standardmäßig auf `all` gesetzt, wenn kein Name angegeben wurde. In der nachstehenden Tabelle werden die Benennungskonventionen für den Plattenpfad beschrieben.

Hinweis – Die Verwendung des globalen Plattenpfadnamens wird dringend empfohlen, weil der globale Plattenpfadname auf dem ganzen Cluster konsistent ist. Der UNIX-Plattenpfadname ist nicht auf dem ganzen Cluster konsistent. Der UNIX-Plattenpfad für eine Platte kann von einem Cluster-Knoten zum anderen unterschiedlich sein. Der Plattenpfad kann auf einem Knoten`c1t0d0` und auf einem anderen Knoten `c2t0d0` sein. Verwenden Sie bei UNIX-Plattenpfadnamen den Befehl `scdidadm -L`, um die UNIX-Plattenpfadnamen den globalen Plattenpfadnamen zuzuordnen, bevor Sie DPM-Befehle ausgeben. Weitere Informationen finden Sie in der Online-Dokumentation `scdidadm`(1M).

TABELLE 3–3 Beispiele für Plattenpfadnamen

Namenstyp	Beispiel Plattenpfadname	Beschreibung
Globaler Plattenpfad	`schost-1:/dev/did/dsk/d1`	Plattenpfad `d1` auf dem Knoten `schost-1`
	`all:d1`	Plattenpfad `d1` auf allen Cluster-Knoten
UNIX-Plattenpfad	`schost-1:/dev/rdsk/c0t0d0s0`	Plattenpfad `c0t0d0s0` auf dem Knoten `schost-1`
	`schost-1:all`	Alle Plattenpfade auf dem Knoten `schost-1`
Alle Plattenpfade	`all:all`	Alle Plattenpfade auf allen Cluster-Knoten

Verwenden von SunPlex-Manager zur Plattenpfadüberwachung

Mit SunPlex-Manager können Sie folgende grundlegende DPM-Verwaltungsaufgaben durchführen:

- Überwachen eines Plattenpfades,
- Aufheben der Überwachung eines Plattenpfades,
- Anzeigen des Status aller Plattenpfade im Cluster.

Verfahrenstechnische Informationen zur Plattenpfadverwaltung mit SunPlex-Manager finden Sie in der Online-Hilfe zu SunPlex-Manager.

Quorum und Quorum-Geräte

Die Cluster-Knoten nutzen gemeinsam Daten und Ressourcen. Daher ist es wichtig, dass ein Cluster niemals in getrennte Partitionen zerfällt, die gleichzeitig aktiv sind. Der CMM stellt selbst bei partitionierten Cluster-Interconnects sicher, dass maximal jeweils ein Cluster in Betrieb ist.

Es gibt zwei Arten von Problemen, die aufgrund der Partitionierung von Clustern auftreten: Split Brain und Amnesie. Zum Split Brain kommt es, wenn der Cluster-Interconnect zwischen den Knoten verloren geht und der Cluster in Teil-Cluster zerfällt, die sich jeweils als die einzige Partition wahrnehmen. Die Ursache sind Kommunikationsprobleme zwischen den Cluster-Knoten. Zur Amnesie kommt es, wenn der Cluster nach dem Herunterfahren mit Cluster-Daten neu startet, die älter als die Daten zum Zeitpunkt des Herunterfahrens sind. Dazu kann es kommen, wenn mehrere Versionen der Framework-Daten auf der Platte gespeichert sind und eine neue Cluster-Zusammensetzung ohne die neueste Version gestartet wird.

Split Brain und Amnesie können vermieden werden, indem jeder Knoten eine Stimme erhält und eine Stimmenmehrzahl für den Betrieb eines Clusters vorgeschrieben wird. Eine Partition mit der Mehrheit der Stimmen hat ein *Quorum* und erhält die Erlaubnis zum Arbeiten. Dieser Mechanismus der Stimmenmehrheit arbeitet einwandfrei, wenn der Cluster aus mehr als zwei Knoten besteht. In einem Zwei-Knoten-Cluster ist zwei eine Mehrheit. Wenn ein solcher Cluster in Partitionen zerfällt, wird eine externe Stimme benötigt, damit eine der Partitionen das Quorum erreicht. Diese externe Stimme wird von einem *Quorum-Gerät* beigesteuert. Als Quorum-Gerät kann jede Platte fungieren, die von beiden Knoten gemeinsam genutzt wird. Als Quorum-Geräte verwendete Platten können Benutzerdaten enthalten.

In Tabelle 3–4 wird beschrieben, wie die Sun Cluster-Software das Quorum zur Vermeidung von Split Brain und Amnesie einsetzt.

TABELLE 3–4 Cluster-Quorum und Split Brain- und Amnesie-Probleme

Partitionstyp	Quorum-Lösung
Split Brain	Erlaubt nur der Partition (dem Teil-Cluster) mit der Stimmenmehrheit, als Cluster zu arbeiten (in dem höchstens eine Partition mit einer solchen Mehrheit vorkommen kann). Sobald ein Knoten den Kampf um das Quorum verloren hat, gerät dieser Knoten in Panik.
Amnesie	Stellt sicher, dass beim Booten eines Clusters mindestens ein Knoten zum Cluster gehört, der Mitglied der letzten Cluster-Mitgliedschaft war (und somit über die neuesten Konfigurationsdaten verfügt).

Der Quorum-Algorithmus arbeitet dynamisch: Da seine Berechnungen durch Cluster-Ereignisse ausgelöst werden, können sich die Rechenergebnisse während der Lebenszeit eines Clusters ändern.

Stimmenanzahl Quorum

Sowohl Cluster-Knoten als auch Quorum-Geräte geben eine Stimme für das Quorum ab. Standardmäßig erhalten Cluster-Knoten eine Stimmenanzahl von Eins für das Quorum, sobald sie booten und Cluster-Mitglieder werden. Knoten können auch eine Stimmenanzahl von Null haben, wenn zum Beispiel der Knoten gerade installiert wird oder wenn ein Verwalter den Knoten in Wartungszustand versetzt hat.

Quorum-Geräte erhalten eine Stimmenanzahl für das Quorum, die sich nach der Anzahl von Knotenverbindungen mit dem Gerät richtet. Wenn ein Quorum-Gerät konfiguriert wird, erhält es eine maximale Stimmenanzahl von *N*-1, wobei *N* der Anzahl der mit dem Quorum-Gerät verbundenen Stimmen entspricht. Ein Quorum-Gerät, das zum Beispiel mit zwei Knoten mit einer Stimmenanzahl von nicht Null verbunden ist, hat einen Quorum-Zählwert von Eins (Zwei minus Eins).

Sie konfigurieren Quorum-Geräte entweder im Verlauf der Cluster-Installation oder später mit den Verfahren, die im *Sun Cluster 3.1 Handbuch Systemverwaltung* beschrieben sind.

Hinweis – Ein Quorum-Gerät trägt nur dann zur Stimmenanzahl bei, wenn mindestens einer der Knoten, mit dem es derzeit verbunden ist, ein Cluster-Mitglied ist. Auch beim Booten von Clustern trägt ein Quorum-Gerät nur dann zur Anzahl bei, wenn mindestens einer der Knoten, mit denen es derzeit verbunden ist, bootet und Mitglied des vor dem Herunterfahren zuletzt gebooteten Clusters war.

Quorum-Konfigurationen

Quorum-Konfigurationen hängen von der Anzahl der Knoten im Cluster ab:

- **Zwei-Knoten-Cluster –** Zwei Quorum-Stimmen sind erforderlich, damit ein Zwei-Knoten-Cluster gebildet werden kann. Diese beiden Stimmen können von den beiden Cluster-Knoten kommen oder von nur einem der Knoten und einem Quorum-Gerät. Trotzdem muss ein Quorum-Gerät für einen Zwei-Knoten-Cluster konfiguriert sein, um sicherzustellen, dass ein einziger Knoten weiterarbeiten kann, wenn der andere Knoten ausfällt.
- **Cluster mit mehr als zwei Knoten** – Sie sollten für jedes Knotenpaar mit gemeinsamem Zugriff auf ein Plattenspeichergehäuse ein Quorum-Gerät angeben. Angenommen, Sie haben einen Drei-Knoten-Cluster, der mit dem Cluster in Abbildung 3–3 vergleichbar ist. In dieser Abbildung haben KnotenA und KnotenB gemeinsam Zugriff auf dasselbe Plattengehäuse, und KnotenB und KnotenC haben gemeinsam Zugriff auf ein anderes Plattengehäuse. Damit läge die Gesamtanzahl von Quorum-Stimmen bei fünf. Drei stammen von den Knoten und zwei von den Quorum-Geräten, die von den Knoten geteilt werden. Ein Cluster benötigt eine Mehrheit von Quorum-Stimmen, um gebildet zu werden.

Die Angabe eines Quorum-Geräts zwischen jedem Knotenpaar, das gemeinsam auf ein Plattenspeichergehäuse zugreift, ist nicht erforderlich und ist von der Sun Cluster-Software nicht zwingend vorgeschrieben. Allerdings kann dies erforderliche Quorum-Stimmen liefern, wenn eine N+1-Konfiguration zu einer Zwei-Cluster-Konfiguration degeneriert und dann der Knoten mit Zugriff auf beide Plattengehäuse ebenfalls ausfällt. Wenn Sie Quorum-Geräte zwischen allen Paaren konfiguriert haben, kann der übrige Knoten weiterhin als Cluster arbeiten.

Beispiele zu diesen Konfigurationen finden Sie in Abbildung 3–3.

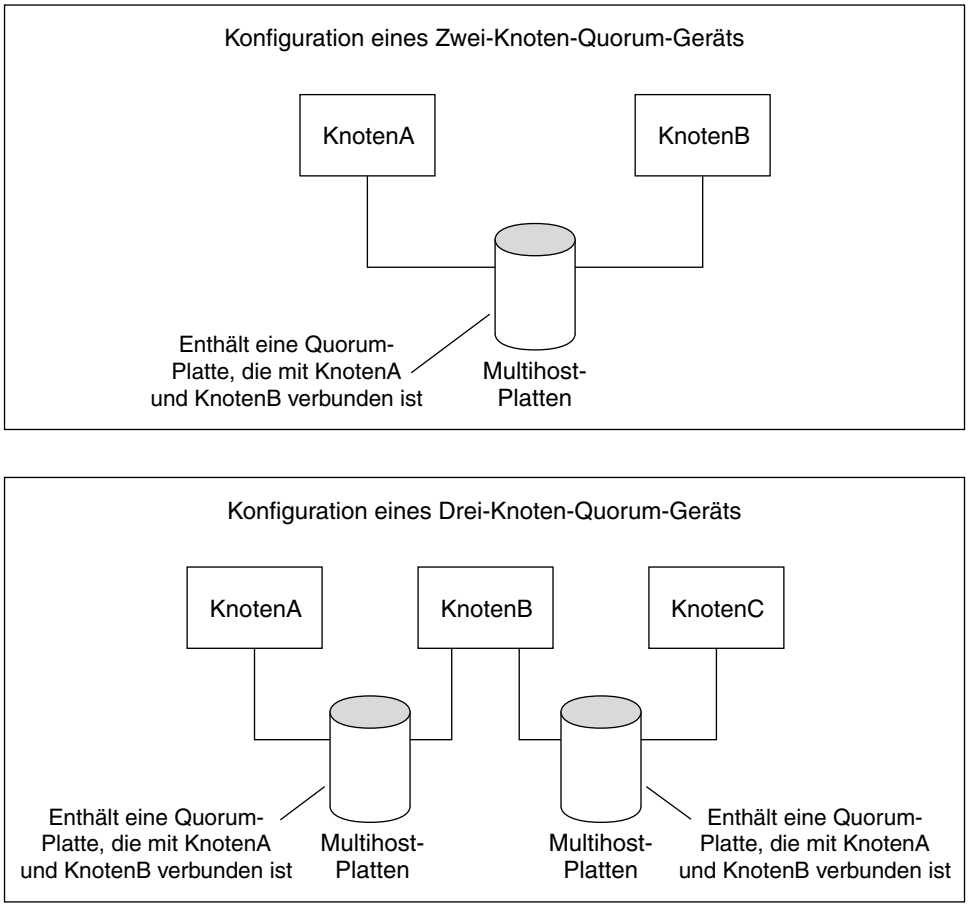


ABBILDUNG 3–3 Beispiele zu Quorum-Gerätekonfigurationen

Quorum-Richtlinien

Befolgen Sie beim Konfigurieren von Quorum-Geräten folgende Richtlinien:

- Legen Sie ein Quorum-Gerät für alle Knoten fest, die an dasselbe gemeinsame Plattenspeichergehäuse angeschlossen sind. Fügen Sie eine Platte als Quorum-Gerät innerhalb des gemeinsam genutzten Gehäuses hinzu, um sicherzustellen, dass bei Versagen eines Knotens die restlichen Knoten das Quorum erhalten und als Master der Plattengerätegruppe im gemeinsamen Gehäuse fungieren können.
- Sie müssen das Quorum-Gerät mit mindestens zwei Knoten verbinden.
- Ein Quorum-Gerät kann jede beliebige SCSI-2- oder SCSI-3-Platte sein, die als Quorum-Gerät mit zwei Ports eingesetzt wird. Platten, die mit mehr als zwei Knoten verbunden sind, müssen SCSI-3-PGR (Persistent Group Reservation) unterstützen, unabhängig davon, ob die Platte als Quorum-Gerät eingesetzt wird. Weitere Informationen finden Sie im Kapitel zur Planung im *Sun Cluster 3.1 Handbuch Softwareinstallation*.
- Sie können eine Platte mit Benutzerdaten als Quorum-Gerät verwenden.

Fehlerschutz

Ein wichtiges Thema bei Clustern ist ein Fehler, der zur Partitionierung des Clusters führt (als *Split Brain* bezeichnet). In diesem Fall können nicht mehr alle Knoten miteinander kommunizieren, so dass einzelne Knoten oder Knoten-Teilsätze ggf. versuchen, Einzel- oder Untermengen-Cluster zu bilden. Jede Untermenge oder Partition kann davon überzeugt sein, alleinigen Zugriff auf die Multihostplatten und die Eigentümerschaft zu haben. Der Versuch mehrerer Knoten, auf die Platten zu schreiben, kann zur Beschädigung von Daten führen.

Der Fehlerschutz schränkt den Knotenzugriff auf die Multihostplatten ein, indem der Zugriff auf die Platten real verhindert wird. Wenn ein Knoten den Cluster verlässt (aufgrund eines Ausfalls oder Partitionierung), wird mit dem Fehlerschutz sichergestellt, dass der Knoten keinen Zugriff mehr auf die Platte hat. Nur aktuelle Mitgliederknoten haben Zugriff auf die Platten. Das sichert die Datenintegrität.

Plattengerätedienste stellen Failover-Funktionen für Dienste zur Verfügung, die Multihostplatten verwenden. Wenn ein aktuell als Primärknoten (Eigentümer) der Plattengruppe fungierendes Cluster-Mitglied ausfällt oder nicht mehr erreicht werden kann, wird ein neuer Primärknoten ausgewählt, und nach einer einer unbedeutenden Unterbrechung kann wieder auf die Plattengerätegruppe zugegriffen werden. Während dieses Prozesses muss der alte Primärknoten den Zugriff auf die Geräte abgeben, bevor der neue Primärknoten gestartet werden kann. Wenn ein Mitglied jedoch aus dem Cluster ausscheidet und nicht mehr erreichbar ist, kann der Cluster diesen Knoten nicht zur Freigabe der Geräte auffordern, für die er als Primärknoten fungierte. Sie brauchen also ein Mittel, mit dessen Hilfe die funktionsfähigen Mitglieder die Steuerung und den Zugriff auf die globalen Geräte der ausgefallenen Mitglieder übernehmen können.

Das SunPlex-System verwendet SCSI-Plattenreservierungen zur Implementierung des Fehlerschutzes. Mit den SCSI-Reservierungen werden die Multihostplatten vor den ausgefallenen Knoten “geschützt” und der Zugriff auf diese Platten wird verhindert.

SCSI-2-Plattenreservierungen unterstützen eine Form der Reservierung, die entweder allen mit der Platte verbundenen Knoten Zugriff erteilt (wenn keine Reservierung vorliegt) oder den Zugriff auf einen einzigen Knoten beschränkt (auf den Knoten, für den die Reservierung gilt).

Wenn ein Cluster-Mitglied erkennt, dass ein anderer Knoten nicht mehr über den Cluster-Interconnect kommuniziert, leitet er ein Fehlerschutzverfahren ein, um den anderen Knoten am Zugriff auf die gemeinsam genutzten Platten zu hindern. Bei einem solchen Fehlerschutz ist es normal, dass der geschützte Knoten in Panik gerät und mit einer Meldung "reservation conflict" in der Konsole reagiert.

Der Reservierungskonflikt tritt nach der Feststellung auf, dass ein Knoten kein Cluster-Mitglied mehr ist, da auf allen Platten, die dieser Knoten mit den restlichen Knoten teilt, eine SCSI-Reservierung erfolgte. Der geschützte Knoten nimmt den Schutz möglicherweise nicht wahr. Wenn er versucht, auf eine der gemeinsam genutzten Platten zuzugreifen, erkennt er die Reservierung und gerät in Panik.

Failfast-Mechanismus zum Fehlerschutz

Der Mechanismus, mit dem Cluster Framework sicherstellt, dass ein ausgefallener Knoten nicht neu booten und in gemeinsam genutzte Speicher schreiben kann, wird als *Failfast* bezeichnet.

Knoten, die Cluster-Mitglieder sind, aktivieren kontinuierlich ein spezifisches ioctl, MHIOCENFAILFAST, für die Platten, auf die sie zugreifen. Hierzu gehören auch die Quorum-Platten. Dieses ioctl ist eine Anweisung für den Plattentreiber. Damit kann sich der Knoten selbst in einen Panik-Zustand versetzen, wenn er nicht auf die Platte zugreifen kann, weil diese von anderen Knoten reserviert wurde.

Das MHIOCENFAILFAST-ioctl löst eine Prüfung der Fehlerrückgaben aus jedem Lese- und Schreibvorgang aus, die von einem Knoten für den Fehlercode `Reservation_Conflict` an die Platte zurückgegeben werden. Das ioctl führt im Hintergrund regelmäßige Testvorgänge auf der Platte aus, um sie auf `Reservation_Conflict` zu prüfen. Sowohl der Kontrollflusspfad im Vordergrund als auch der im Hintergrund geraten in Panik, wenn `Reservation_Conflict` zurückgegeben wird.

Bei SCSI-2-Platten sind die Reservierungen nicht dauerhaft — sie werden beim erneuten Booten von Knoten gelöscht. Bei SCSI-3-Platten mit PGR (Persistent Group Reservation) werden die Reservierungsinformationen auf der Platte gespeichert und bleiben auch nach dem Booten von Knoten erhalten. Der Failfast-Mechanismus arbeitet immer gleich, unabhängig davon, ob Sie SCSI-2- oder SCSI-3-Platten verwenden.

Wenn ein Knoten die Konnektivität mit anderen Knoten im Cluster verliert und nicht zu einer Partition gehört, die ein Quorum erzielen kann, wird er erzwungenermaßen von einem anderen Knoten aus dem Cluster entfernt. Ein anderer Knoten führt als Teil der Partition, die ein Quorum erzielt, Reservierungen auf den gemeinsam genutzten

Platten aus. Wenn der Knoten ohne Quorum nun versucht, auf die gemeinsam genutzten Platten zuzugreifen, erhält er einen Reservierungskonflikt als Antwort und gerät infolge des Failfast-Mechanismus in Panik.

Nach der Panik kann der Knoten neu booten und versuchen, dem Cluster wieder beizutreten oder am OpenBoot PROM (OBP) zu bleiben. Welche Aktion eingeleitet wird, bestimmt die Einstellung des `auto-boot?`-Parameters im OBP.

Datenträger-Manager

Das SunPlex-System setzt Software für die Datenträgerverwaltung ein, um die Datenverfügbarkeit durch die Verwendung von Spiegelungs- und Hot-Spare-Platten zu verbessern und Plattenversagen sowie Ersetzungsvorgänge abzuwickeln.

Das SunPlex-System hat keine eigene interne Datenträger-Manager-Komponente, sondern arbeitet mit den folgenden Datenträger-Managern:

- Solaris Volume Manager
- VERITAS Volume Manager

Die Datenträgerverwaltungs-Software im Cluster unterstützt Folgendes:

- Failover-Abwicklung bei Knotenversagen,
- Multipath-Unterstützung von unterschiedlichen Knoten aus,
- Transparenter Remote-Zugriff auf Plattengerätegruppen.

Sobald Datenträgerverwaltungs-Objekte der Steuerung des Clusters unterstellt sind, werden sie zu Plattengerätegruppen. Informationen zu Datenträger-Manager finden Sie in der Dokumentation zur Datenträger-Manager-Software.

Hinweis – Ein wichtiger Gesichtspunkt bei der Planung Ihrer Plattensätze oder Plattengruppen ist das Verständnis, wie die ihnen zugeordneten Plattengerätegruppen den Anwendungsressourcen (Daten) innerhalb des Clusters zugeordnet sind. Erläuterungen zu diesen Fragen finden Sie im Sun Cluster 3.1 Handbuch Softwareinstallation und im Sun Cluster 3.1 Data Services Installation and Configuration Guide.

Datendienste

Der Begriff *Datendienst* beschreibt eine Anwendung von Drittherstellern wie Oracle oder Sun ONE Web Server, die eher zur Ausführung auf einem Cluster als auf einem Einzelserver konfiguriert wurde. Ein Datendienst besteht aus einer Anwendung, spezialisierten Sun Cluster-Konfigurationsdateien und Sun Cluster-Verwaltungsmethoden, mit denen die nachfolgenden Anwendungsaktionen gesteuert werden.

- Starten
- Anhalten
- Überwachen und Ergreifen von Korrekturmaßnahmen

In Abbildung 3–4 wird eine Anwendung, die auf einem einzigen Anwendungsserver ausgeführt wird (Einzelservermodell) mit derselben Anwendung auf einem Cluster (Cluster-Servermodell) verglichen. Beachten Sie, dass aus der Benutzerperspektive kein Unterschied zwischen den beiden Konfigurationen besteht. Die Cluster-Anwendung wird ggf. schneller ausgeführt und zeigt eine bessere Hochverfügbarkeit.

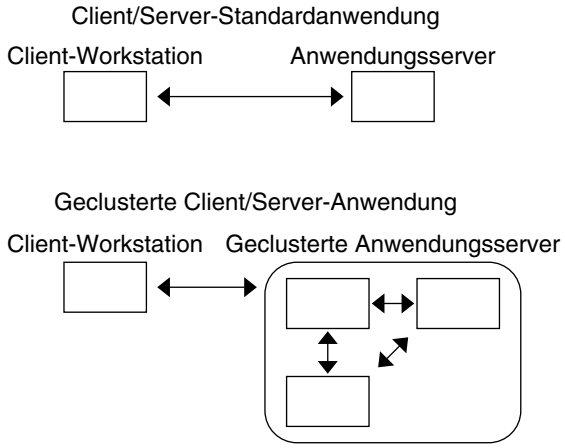


ABBILDUNG 3–4 Standardkonfiguration versus Client/Server-Konfiguration als Cluster

Beim Einzelservermodell konfigurieren Sie die Anwendung für den Zugriff auf den Server über eine bestimmte öffentliche Netzwerkschnittstelle (einen Hostnamen). Der Hostname ist diesem realen Server zugeordnet.

Beim geclusterten Servermodell ist die öffentliche Netzwerkschnittstelle ein *logischer Hostname* oder eine *gemeinsam genutzte Adresse*. Der Begriff *Netzwerkressourcen* bezeichnet sowohl die logischen Hostnamen als auch die gemeinsam genutzten Adressen.

Für bestimmte Datendienste müssen Sie entweder logische Hostnamen oder gemeinsam genutzte Adressen als Netzwerkschnittstellen angeben — sie sind nicht austauschbar. Für andere Datendienste können Sie wahlweise logische Hostnamen oder gemeinsam genutzte Adressen angeben. Details zum jeweils erforderlichen Schnittstellentyp finden Sie in den Angaben zur Installation und Konfiguration für den jeweiligen Datendienst.

Eine Netzwerkressource ist keinem spezifischen realen Server zugeordnet — sie kann zwischen den realen Servern migrieren.

Eine Netzwerkressource ist zunächst einem Knoten zugeordnet, dem *Primärknoten*. Wenn der Primärknoten ausfällt, werden Netzwerkressource und Anwendungsressource mit einem Failover-Vorgang auf einen anderen Cluster-Knoten umgeleitet (einen Sekundärknoten). Wenn die Netzwerkressource ein Failover durchführt, wird die Anwendungsressource nach einer kurzen Verzögerung auf dem Sekundärknoten weiter ausgeführt.

In Abbildung 3–5 wird das Einzelservermodell mit dem geclusterten Servermodell verglichen. Beachten Sie, dass eine Netzwerkressource (in diesem Beispiel ein logischer Hostname) in einem geclusterten Servermodell zwischen mindestens zwei Cluster-Knoten wechseln kann. Die Anwendung ist zur Verwendung dieses logischen Hostnamens anstelle eines Hostnamens konfiguriert, der einem bestimmten Server zugeordnet ist.

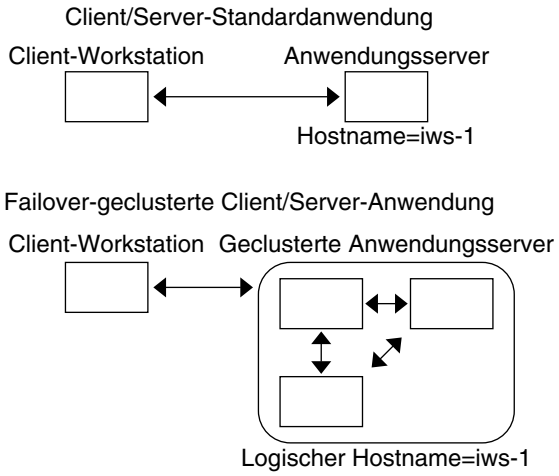


ABBILDUNG 3–5 Festgelegter Hostname versus logischer Hostname

Eine gemeinsam genutzte Adresse ist zunächst ebenfalls einem Knoten zugeordnet. Dieser Knoten wird als globaler Schnittstellenknoten (GIF-Knoten) bezeichnet. Eine gemeinsam genutzte Adresse wird als einzige Netzwerkschnittstelle zum Cluster eingesetzt. Sie wird als *globale Schnittstelle* bezeichnet.

Der Unterschied zwischen dem Modell logischer Hostname und dem Modell Scalable-Dienst besteht darin, dass beim letztgenannten auf jedem Knoten die gemeinsam genutzte Adresse auf der Schleifenschnittstelle ebenfalls aktiv konfiguriert ist. Dank dieser Konfiguration können mehrere Instanzen eines Datendienstes auf mehreren Knoten gleichzeitig aktiv sein. Der Begriff “Scalable-Dienst” bedeutet, dass Sie die CPU-Leistung für die Anwendung durch Hinzufügen zusätzlicher Cluster-Knoten erhöhen können und die Leistung verbessert wird.

Wenn ein GIF-Knoten ausfällt, kann die gemeinsam genutzte Adresse auf einen anderen Knoten verlegt werden, auf dem ebenfalls eine Instanz der Anwendung ausgeführt wird (wobei dieser andere Knoten zum GIF-Knoten wird). Oder die gemeinsam genutzte Adresse wird mit einem Failover auf einen anderen Cluster-Knoten verschoben, auf dem die Anwendung zuvor nicht ausgeführt wurde.

In Abbildung 3–6 wird die Einzelserverkonfiguration mit der Cluster-Konfiguration mit Scalable-Diensten verglichen. Beachten Sie, dass die gemeinsam genutzte Adresse in der Scalable-Dienst-Konfiguration auf allen Knoten vorhanden ist. Ähnlich wie bei der Verwendung eines logischen Hostnamens für Failover-Datendienste wird die Anwendung zur Verwendung dieser gemeinsam genutzten Adresse anstelle eines einem bestimmten Server zugeordneten Hostnamens konfiguriert.

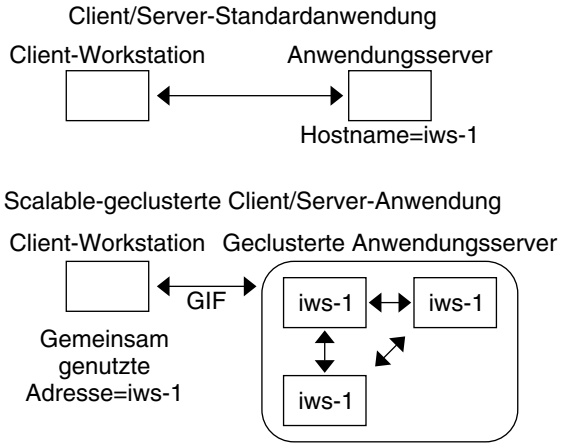


ABBILDUNG 3–6 Festgelegter Hostname versus gemeinsam genutzte Adresse

Datendienstmethoden

Die Sun Cluster-Software stellt eine Reihe von Dienstverwaltungsmethoden zur Verfügung. Diese Methoden werden vom Ressourcengruppen-Manager (RGM) gesteuert und eingesetzt, um die Anwendung auf den Cluster-Knoten zu starten, anzuhalten und zu überwachen. Mit diesen Methoden, der Cluster-Framework-Software und den Multihostplatten können die Anwendungen als Failover- oder Scalable-Datendienste verwendet werden.

Der RGM verwaltet auch Ressourcen im Cluster einschließlich der Anwendungsinstanzen und Netzwerkressourcen (logische Hostnamen und gemeinsam genutzte Adressen).

Zusätzlich zu den Methoden der Sun Cluster-Software stellt das SunPlex-System auch eine API und verschiedene Datendienst-Entwicklungstools zur Verfügung. Mit diesen Tools können Anwendungsprogrammierer die Datendienstmethoden entwickeln, die sie zum Ausführen anderer Anwendungen als hoch verfügbare Datendienste mit der Sun Cluster-Software benötigen.

Failover-Datendienste

Wenn der Knoten ausfällt, auf dem der Datendienst ausgeführt wird (der Primärknoten), migriert der Dienst ohne Benutzereingriff auf einen anderen funktionsfähigen Knoten. Failover-Dienste verwenden eine *Failover-Ressourcengruppe*, die Ressourcen für Anwendungsinstanzen und Netzwerkressourcen enthält (*logische Hostnamen*). Logische Hostnamen sind IP-Adressen, die auf einem Knoten als aktiv konfiguriert werden können. Später werden sie auf dem Originalknoten automatisch als inaktiv und auf einem anderen Knoten als aktiv konfiguriert.

Für Failover-Datendienste werden die Anwendungsinstanzen nur auf einem einzigen Knoten ausgeführt. Wenn der Fehler-Monitor einen Fehler erkennt, versucht er entweder, die Instanz auf demselben Knoten erneut zu starten oder die Instanz auf einem anderen Knoten zu starten (Failover), je nachdem, wie der Datendienst konfiguriert wurde.

Scalable-Datendienste

Der Scalable-Datendienst kann aktive Instanzen auf mehreren Knoten ausführen. Scalable-Dienste verwenden zwei Ressourcengruppen: eine *Scalable-Ressourcengruppe* mit den Anwendungsressourcen und eine Failover-Ressourcengruppe mit den Netzwerkressourcen (*gemeinsam genutzte Adressen*), von denen der Scalable-Dienst abhängt. Die Scalable-Ressourcengruppe kann auf mehreren Knoten online sein, so dass mehrere Instanzen dieses Dienstes gleichzeitig ausgeführt werden können. Die Failover-Ressourcengruppe, welche die gemeinsam genutzten Adressen hostet, ist jeweils nur auf einem Knoten online. Alle Knoten, die einen Scalable-Dienst hosten, verwenden dieselbe gemeinsam genutzte Adresse, um den Dienst zu hosten.

Dienstanforderungen erreichen den Cluster über eine einzige Netzwerkschnittstelle (die globale Schnittstelle) und werden anhand unterschiedlicher, vordefinierter Algorithmen auf die Knoten verteilt, die im Rahmen des *Lastausgleichsverfahrens* eingestellt wurden. Der Cluster kann die Dienstlast mithilfe des Lastausgleichsverfahrens zwischen mehreren Knoten ausgleichen. Beachten Sie, dass mehrere globale Schnittstellen auf anderen Knoten weitere, gemeinsam genutzte Adressen hosten können.

Bei Scalable-Diensten werden die Anwendungsinstanzen auf mehreren Knoten gleichzeitig ausgeführt. Wenn der Knoten mit der globalen Schnittstelle ausfällt, wird die globale Schnittstelle auf einen anderen Knoten verschoben. Wenn eine Anwendungsinstanz fehlschlägt, versucht die Instanz, auf demselben Knoten neu zu starten.

Wenn eine Anwendungsinstanz nicht auf demselben Knoten neu gestartet werden kann und ein anderer, nicht verwendeter Knoten für die Ausführung des Dienstes konfiguriert ist, wird der Dienst im Rahmen eines Failover-Vorgangs auf den nicht verwendeten Knoten verschoben. Andernfalls wird er weiter auf den restlichen Knoten ausgeführt und führt möglicherweise zu einer Reduzierung des Dienstdurchsatzes.

Hinweis – Der TCP-Zustand für jede Anwendungsinstanz bleibt auf dem Knoten mit der Instanz, nicht auf dem Knoten mit der globalen Schnittstelle. Deswegen hat ein Ausfall des Knotens mit der globalen Schnittstelle keine Auswirkung auf die Verbindung.

Abbildung 3–7 zeigt ein Beispiel für eine Failover- und eine Scalabe-Ressourcengruppe und die Abhängigkeiten für Scalable-Dienste. Dieses Beispiel enthält drei Ressourcengruppen. Die Failover-Ressourcengruppe enthält Anwendungsressourcen für hoch verfügbaren DNS und Netzwerkressourcen, die sowohl vom hoch verfügbaren DNS als auch vom hoch verfügbaren Apache Web Server genutzt werden. Die Scalable-Ressourcengruppen enthalten nur Anwendungsinstanzen des Apache Web Servers. Beachten Sie, dass Ressourcengruppenabhängigkeiten zwischen den Scalable- und den Failover-Ressourcengruppen (durchgezogene Linien) bestehen, und dass alle Apache-Anwendungsressourcen von der Netzwerkressource `schost-2` abhängen, die eine gemeinsam genutzte Adresse ist (gestrichelte Linien).

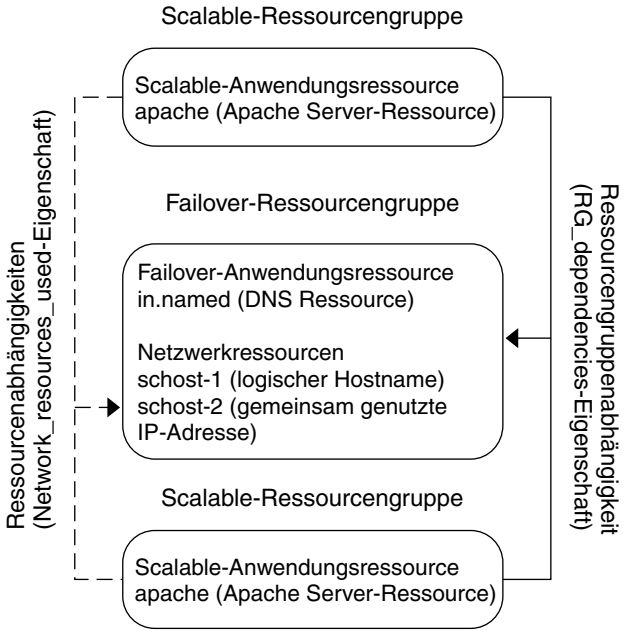


ABBILDUNG 3–7 Beispiel für Failover- und Scalable-Ressourcengruppen

Scalable-Dienst-Architektur

Das wichtigste Ziel eines Cluster-Netzwerks besteht darin, Skalierbarkeit für Datendienste zu ermöglichen. Skalierbarkeit bedeutet, dass ein Datendienst trotz steigender Belastung eine konstante Antwortzeit beibehalten kann, weil dem Cluster neue Knoten hinzugefügt und neue Serverinstanzen ausgeführt werden. Solche Dienste werden als Scalable-Datendienste bezeichnet. Ein gutes Beispiel für einen Scalable-Datendienst ist ein Webdienst. In der Regel besteht ein Scalable-Datendienst aus mehreren Instanzen, von denen jede auf unterschiedlichen Cluster-Knoten ausgeführt wird. Insgesamt verhalten sich diese Instanzen aus Sicht eines Remote-Clients wie ein einziger Dienst und führen die Funktionen dieses Dienstes aus. Ein Scalable-Webdienst kann zum Beispiel aus mehreren `httpd`-Dämonen auf unterschiedlichen Knoten bestehen. Jeder `httpd`-Dämon kann eine Client-Anforderung bedienen. Welcher Dämon die Anforderung bedient, hängt vom *Lastausgleichsverfahren* ab. Die Antwort an den Client kommt scheinbar vom Dienst und nicht vom konkreten Dämon, der die Anforderung bedient hat, so dass der Eindruck eines einzigen Dienstes gewahrt wird.

Ein Scalable-Dienst besteht aus Folgendem:

- die Scalable-Dienste unterstützende Netzwerkinfrastruktur,
- Lastausgleichsverfahren,

- Unterstützung für Netzwerk und Datendienste (unter Verwendung des Ressourcengruppen-Manager).

Die nachstehende Abbildung zeigt eine Scalable-Dienst-Architektur.

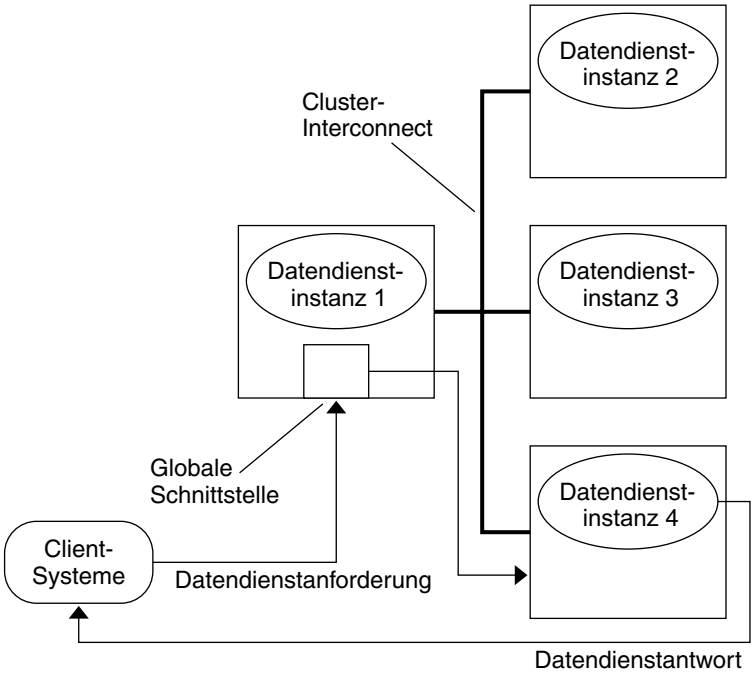


ABBILDUNG 3–8 Scalable-Dienst-Architektur

Die Knoten, welche die globale Schnittstelle nicht hosten (Proxy-Knoten), hosten die gemeinsam genutzte Adresse auf ihren Schleifenschnittstellen. Die an der globalen Schnittstelle eintreffenden Pakete werden anhand konfigurierbarer Lastausgleichsverfahren auf andere Cluster-Knoten verteilt. Die möglichen Lastausgleichsverfahren werden im Folgenden beschrieben.

Lastausgleichsverfahren

Der Lastausgleich verbessert die Leistung der Scalable-Dienste sowohl bei der Antwortzeit als auch beim Durchsatz.

Es gibt zwei Klassen von Scalable-Datendiensten: *reine* und *sticky*. Ein reiner Dienst ist ein Dienst, bei dem jede Instanz Client-Anforderungen beantworten kann. Ein Sticky-Dienst ist ein Dienst, bei dem ein Client Anforderungen an dieselbe Instanz sendet. Diese Anforderungen werden nicht an andere Instanzen umgeleitet.

Ein reiner Dienst verwendet ein gewichtetes Lastausgleichsverfahren. Bei diesem Lastausgleichsverfahren werden Client-Anforderungen standardmäßig gleichmäßig auf die Serverinstanzen im Cluster verteilt. Angenommen, dass jeder Knoten in einem Drei-Knoten-Cluster eine Gewichtung von 1 hat. Jeder Knoten bedient 1/3 der Anforderungen aller beliebigen Clients im Auftrag des Dienstes. Gewichtungen können jederzeit vom Verwalter über die `scrgadm(1M)`-Befehlsschnittstelle oder über die GUI von SunPlex-Manager geändert werden.

Ein Sticky-Dienst hat zwei Typen, *normal-sticky* und *Platzhalter-Sticky*. Mit Sticky-Diensten können Sitzungen auf Anwendungsebene gleichzeitig über mehrere TCP-Verbindungen ausgeführt werden, um den Zustandsspeicher (Anwendungssitzungszustand) gemeinsam zu nutzen.

Mit normalen Sticky-Diensten kann ein Client den Zustand zwischen mehreren gleichzeitigen TCP-Verbindungen teilen. Der Client wird als “sticky” bezeichnet, weil die Serverinstanz einen einzigen Port abhört. Der Client hat die Sicherheit, dass alle seine Anforderungen an dieselbe Serverinstanz gehen. Voraussetzung hierfür ist, dass die Instanz aktiv und zugänglich bleibt und das Lastausgleichsverfahren nicht geändert wird, solange der Dienst online ist.

Ein Webbrowser auf dem Client stellt zum Beispiel mithilfe von drei verschiedenen TCP-Verbindungen eine Verbindung mit einer gemeinsam genutzten IP-Adresse an Port 80 her, aber die Verbindungen tauschen zwischengespeicherte Sitzungsinformationen mit dem Dienst aus.

Eine Verallgemeinerung eines Sticky-Verfahrens erstreckt sich auf mehrere Scalable-Dienste, die im Hintergrund Sitzungsinformationen bei derselben Instanz austauschen. Wenn diese Dienste im Hintergrund Sitzungsinformation bei derselben Instanz austauschen, wird der Client als “sticky” bezeichnet, weil mehrere Serverinstanzen auf demselben Knoten unterschiedliche Ports abhören.

Ein E-Commerce-Kunde füllt seinen Einkaufskorb zum Beispiel mit Artikeln und verwendet das normale HTTP an Port 80, wechselt dann jedoch zum Senden sicherer Daten mit SSL auf Port 443, um die Artikel im Einkaufskorb mit Kreditkarte zu bezahlen.

Platzhalter-Sticky-Dienste verwenden dynamisch zugewiesene Port-Nummern, erwarten jedoch trotzdem, dass die Client-Anforderungen an denselben Knoten geleitet werden. Der Client ist in Bezug auf dieselbe IP-Adresse “Sticky-Platzhalter” bei den Ports.

Ein gutes Beispiel dieses Verfahrens ist der passive FTP-Modus. Ein Client stellt an Port 21 eine Verbindung mit einem FTP-Server her und wird vom Server angewiesen, eine neue Verbindung mit einem Listener-Port-Server im dynamischen Port-Bereich herzustellen. Alle Anforderungen an diese IP-Adresse werden an denselben Knoten weitergeleitet, den der Server dem Client mit den Steuerinformationen angegeben hat.

Beachten Sie, dass das gewichtete Lastausgleichsverfahren für jedes dieser Sticky-Verfahren standardmäßig gültig ist, so dass die ursprüngliche Client-Anforderung an die vom Lastausgleich vorgegebene Instanz geleitet wird.

Nachdem der Client eine Affinität zum Knoten mit der ausgeführten Instanz aufgebaut hat, werden zukünftige Anforderungen an diese Instanz geleitet, solange der Knoten zugänglich ist und das Lastausgleichsverfahren nicht geändert wird.

Weitere Details zu spezifischen Lastausgleichsverfahren werden nachstehend erläutert.

- Gewichtet. Die Last wird entsprechend den angegebenenen Gewichtungswerten auf mehrere Knoten verteilt. Dieses Verfahren wird mit dem `LB_WEIGHTED`-Wert für die `Load_balancing_weights`-Eigenschaft eingestellt. Wenn die Gewichtung für einen Knoten nicht ausdrücklich angegeben ist, beträgt die Standardgewichtung für diesen Knoten Eins.

  Das gewichtete Verfahren leitet einen gewissen Prozentsatz des Datenverkehrs von Clients auf einen bestimmten Knoten. Bei X=Gewichtung und A=Gesamtgewichtung aller aktiven Knoten ist davon auszugehen, dass etwa X/A aller neuen Verbindungen zu einem aktiven Knoten geleitet werden, wenn die Gesamtanzahl der Verbindungen groß genug ist. Dieses Verfahren befasst sich nicht mit einzelnen Anforderungen.

  Beachten Sie, dass es sich nicht um ein Round-Robin-Verfahren handelt. Bei einem Round-Robin-Verfahren würde jede Anforderung eines Clients an einen unterschiedlichen Knoten geleitet: die erste Anforderung an Knoten 1, die zweite Anforderung an Knoten 2 und so weiter.
- Sticky. Bei diesem Verfahren ist der Satz an Ports bei der Konfiguration der Anwendungsressourcen bekannt. Dieses Verfahren wird mit dem `LB_STICKY`-Wert der `Load_balancing_policy`-Ressourceneigenschaft eingestellt.
- Sticky-Platzhalter. Dieses Verfahren ist eine Obermenge des normalen "Sticky"-Verfahrens. Bei einem durch die IP-Adresse identifizierten Scalable-Dienst werden die Ports vom Server zugewiesen (und sind vorher nicht bekannt). Die Ports können sich ändern. Dieses Verfahren wird mit dem `LB_STICKY_WILD`Wert der `Load_balancing_policy`-Ressourceneigenschaft eingestellt.

Failback-Einstellungen

Ressourcengruppen wechseln beim Failover von einem Knoten auf einen anderen. In diesem Fall wird der ursprüngliche Sekundärknoten zum neuen Primärknoten. Die Failback-Einstellungen legen die Aktionen fest, die ausgeführt werden, sobald der ursprüngliche Primärknoten wieder online gebracht wird. Die Optionen sind, entweder den ursprünglichen Primärknoten wieder zum Primärknoten zu machen (Failback) oder den aktuellen Primärknoten als solchen zu belassen. Sie geben die gewünschte Option mit der Einstellung der `Failback`-Ressourcengruppeneigenschaft an.

Bei bestimmten Instanzen kann die Failback-Einstellung zu einer reduzierten Verfügbarkeit der Ressourcengruppe führen, wenn der Originalknoten mit der Ressource ausfällt und mehrmals neu bootet.

Fehler-Monitore der Datendienste

Jeder SunPlex-Datendienst stellt einen Fehler-Monitor zur Verfügung, der regelmäßig den Datendienst auf einwandfreies Funktionieren prüft. Ein Fehler-Monitor prüft, ob der bzw. die Anwendungsdämon(en) ausgeführt und die Clients bedient werden. Aufgrund der von den Testsignalen zurückgebenen Informationen können vordefinierte Aktionen wie ein Neustart eines Dämons oder das Einleiten eines Failovers ausgelöst werden.

Entwickeln von neuen Datendiensten

Sun stellt Vorlagen für Konfigurationsdateien und Verwaltungsmethoden zur Verfügung, mit denen Sie unterschiedliche Anwendungen zu Failover- oder Scalable-Diensten innerhalb eines Clusters machen können. Wenn die Anwendung, die Sie als Failover- oder Scalable-Dienst ausführen möchten, von Sun derzeit nicht angeboten wird, können Sie diese Anwendung mit einer API oder mit der DSET-API als Failover- oder Scalable-Dienst konfigurieren.

Es gibt eine Reihe von Kriterien, mit denen Sie feststellen können, ob eine Anwendung als Failover-Dienst ausgeführt werden kann. Die Beschreibung des spezifischen Kriteriums finden Sie in den SunPlex-Dokumenten zu den APIs, die Sie für Ihre Anwendung verwenden können.

Wir stellen Ihnen hier einige Richtlinien vor, mit deren Hilfe Sie erkennen können, ob Ihr Dienst die Scalable-Datendienst-Architektur nutzbringend einsetzen kann. Allgemeine Informationen zu Scalable-Diensten finden Sie unter „Scalable-Datendienste“ auf Seite 63.

Neue Dienste, die folgende Richtlinien erfüllen, können Scalable-Dienste verwenden. Wenn ein vorhandener Dienst diese Richtlinien nicht genau einhält, müssen Teile davon möglicherweise neu geschrieben werden, damit der Dienst den Richtlinien entspricht.

Ein Scalable-Datendienst weist folgende Merkmale auf. Erstens: Ein solcher Dienst besteht aus einer oder mehreren Server*instanz(en)*. Jede Instanz wird auf einem anderen Knoten des Clusters ausgeführt. Zwei oder mehr Instanzen desselben Dienstes können nicht auf demselben Knoten ausgeführt werden.

Zweitens: Wenn der Dienst einen externen logischen Datenspeicher zur Verfügung stellt, muss der gleichzeitige Zugriff auf diesen Speicher durch mehrere Serverinstanzen synchronisiert werden, um den Verlust von Aktualisierungen oder das Lesen von Daten zu verhindern, wenn diese gerade geändert werden. Beachten Sie, dass die Bezeichnung “extern” verwendet wird, um den Speicher vom Arbeitsspeicherstatus zu unterscheiden, und der Begriff “logisch”, weil der Speicher als einzelne Entität angezeigt wird, auch wenn er möglicherweise repliziert ist. Außerdem hat dieser logische Datenspeicher die Eigenschaft, dass eine Aktualisierung des Speichers durch eine beliebige Instanz sofort von anderen Instanzen erkannt wird.

Das SunPlex-System stellt einen solchen externen Speicher über das Cluster-Dateisystem und die globalen Partitionen im raw-Modus zur Verfügung. Angenommen, ein Dienst schreibt neue Daten in eine externe Protokolldatei oder ändert die dort vorhandenen Daten. Wenn mehrere Instanzen dieses Dienstes ausgeführt werden, hat jede Instanz Zugriff auf dieses externe Protokoll und jede kann gleichzeitig auf dieses Protokoll zugreifen. Jede Instanz muss den Zugriff auf dieses Protokoll synchronisieren, andernfalls kommt es zu Interferenzen zwischen den Instanzen. Der Dienst kann die normale Solaris-Dateisperrung über `fcntl(2)` und `lockf(3C)` verwenden, um die gewünschte Synchronisierung zu erreichen.

Ein weiteres Beispiel für diesen Speichertyp ist eine Backend-Datenbank wie das hoch verfügbare Oracle oder Oracle Parallel Server/Real Application Clusters. Beachten Sie, dass bei diesem Typ von Backend-Datenbankserver die Synchronisierung über Datenbankabfragen oder Aktualisierungstransaktionen integriert ist, so dass bei mehreren Serverinstanzen nicht jede einzelne Instanz eine eigene Synchronisierung durchführen muss.

Ein Beispiel für einen nicht skalierbaren Dienst ist die aktuelle Zusammensetzung des IMAP-Servers von Sun. Der Dienst aktualisiert einen Speicher, aber dieser Speicher ist privat. Wenn mehrere IMAP-Instanzen in diesen Speicher schreiben, werden die jeweiligen Einträge überschrieben, weil die Aktualisierungen nicht synchronisiert sind. Der IMAP-Server muss für die Synchronisierung eines gleichzeitigen Zugriffs umgeschrieben werden.

Beachten Sie schließlich, dass Instanzen über private Daten verfügen können, die von den Daten anderer Instanzen getrennt sind. In einem solchen Fall muss der Dienst den gleichzeitigen Zugriff nicht selbst sznchronisieren, weil die Daten privat sind und nur diese Instanz sie bearbeiten kann. In diesem Fall müssen Sie darauf achten, diese privaten Daten nicht unter dem Cluster-Dateisystem zu speichern, weil die Möglichkeit eines globalen Zugriffs besteht.

Datendienst-API und API der Datendienst-Entwicklungsbibliothek

Das SunPlex-System stellt Folgendes zur Verfügung, um Anwendungen hoch verfügbar zu machen:

- Datendienste, die zum Lieferumfang des SunPlex-Systems gehören,
- Eine Datendienst-API,
- Eine API für die Datendienst-Entwicklungsbibliothek,
- Einen “generischen” Datendienst.

Im *Sun Cluster 3.1 Data Services Installation and Configuration Guide* wird die Installation und Konfiguration der Datendienste beschrieben, die zum Lieferumfang des SunPlex-Systems gehören. Im *Sun Cluster 3.1 Entwicklerhandbuch Datendienste* wird das Einrichten anderer Anwendungen als hoch verfügbar im Sun Cluster-Framework beschrieben.

Mit den Sun Cluster-APIs können Anwendungsprogrammierer Fehler-Monitore sowie Skripts zum Starten und Anhalten von Datendienstinstanzen entwickeln. Mit diesen Tools kann eine Anwendung als Failover- oder Scalable-Datendienst eingerichtet werden. Zudem stellt das SunPlex-System einen "generischen" Datendienst zur Verfügung, mit dem die benötigten Start- und Stoppmethoden für eine Anwendung, die als Failover- oder Scalable-Dienst ausgeführt werden soll, schnell generiert werden können.

Verwenden des Cluster-Interconnect für den Datendienstverkehr

Ein Cluster benötigt mehrere Netzwerkverbindungen zwischen den Knoten, die den Cluster-Interconnect bilden. Die Cluster-Software verwendet mehrere Interconnects, um die Hochverfügbarkeit und die Leistung zu verbessern. Für den internen Datenverkehr (zum Beispiel Dateisystemdaten oder Scalable-Dienstdaten) werden Meldungen über alle verfügbaren Interconnects im Round-Robin-Verfahren gesendet.

Der Cluster-Interconnect steht auch Anwendungen zur hoch verfügbaren Kommunikation zwischen den Knoten zur Verfügung. Eine verteilte Anwendung kann zum Beispiel Komponenten auf unterschiedlichen Knoten ausführen, die miteinander kommunizieren müssen. Indem der Cluster-Interconnect anstelle des öffentlichen Netzwerks verwendet wird, bleiben diese Verbindungen beim Versagen einer einzelnen Verknüpfung bestehen.

Um den Cluster-Interconnect für die Datenverbindung zwischen den Knoten zu verwenden, muss eine Anwendung die während der Cluster-Installation konfigurierten privaten Hostnamen verwenden. Wenn zum Beispiel der private Hostname für Knoten 1 `clusternode1-priv` ist, kommunizieren Sie mit diesem Namen über den Cluster-Interconnect mit Knoten 1. TCP-Sockets, die mit diesem Namen geöffnet wurden, werden über den Cluster-Interconnect geleitet und können bei einem Netzwerkfehler transparent umgeleitet werden.

Beachten Sie, dass der Cluster-Interconnnect jeden beliebigen, während der Installation gewählten Namen verwenden kann, da die privaten Hostnamen während der Installation konfiguriert werden können. Den tatsächlichen Namen erhalten Sie über `scha_cluster_get`(3HA) mit dem Argument `scha_privatelink_hostname_node`.

Bei Verwendung des Cluster-Interconnects auf Anwendungsebene wird ein einizger Interconnect zwischen jedem Knotenpaar eingesetzt; wenn möglich werden jedoch getrennte Interconnects für unterschiedliche Knotenpaare eingesetzt. Angenommen, eine Anwendung wird auf drei Knoten ausgeführt und kommuniziert über den Cluster-Interconnect. Die Kommunikation zwischen den Knoten 1 und 2 kann über die Schnittstelle `hme0` erfolgen, während die Kommunikation zwischen den Knoten 1

und 3 über die Schnittstelle `qfe1` erfolgen kann. Das bedeutet, dass die Anwendungskommunikation zwischen zwei beliebigen Knoten auf einen einzigen Interconnect beschränkt ist, während die Cluster-interne Kommunikation über alle Interconnects verteilt wird.

Beachten Sie, dass die Anwendung den Interconnect mit dem internen Cluster-Datenverkehr teilt, so dass die für die Anwendung verfügbare Bandbreite von der Bandbreite abhängt, die für anderen Cluster-Datenverkehr verwendet wird. Im Fall eines Versagens kann der interne Datenverkehr im Round-Robin-Verfahren auf die restlichen Interconnects umgeleitet werden, während die Anwendungsverbindungen mit einem versagenden Interconnect auf einen funktionierenden Interconnect umgeschaltet werden können.

Zwei Adresstypen unterstützen den Cluster-Interconnect, und mit `gethostbyname`(3N) werden für einen privaten Hostnamen in der Regel zwei IP-Adressen zurückgegeben. Die erste Adresse wird als *logische Pro-Paar-Adresse* bezeichnet, die zweite Adresse als *logische Pro-Knoten-Adresse*.

Jedem Knotenpaar wird eine eigene logische Pro-Paar-Adresse zugewiesen. Dieses kleine logische Netzwerk unterstützt Failover-Vorgänge von Verbindungen. Jedem Knoten wird außerdem eine feste Pro-Knoten-Adresse zugewiesen. Mit anderen Worten, die logischen Pro-Paar-Adressen für `clusternode1-priv` sind auf jedem Knoten anders, während die logische Pro-Knoten-Adresse für `clusternode1-priv` für jeden Knoten gleich ist. Ein Knoten hat jedoch für sich selbst keine Pro-Paar-Adresse, so dass mit `gethostbyname` (`clusternode1-priv`) auf Knoten 1 nur die logische Pro-Knoten-Adresse zurückgegeben wird.

Beachten Sie, dass Anwendungen, die Verbindungen über den Cluster-Interconnect zulassen und dann aus Sicherheitsgründen die IP-Adresse überprüfen, alle IP-Adressen überprüfen müssen, die mit `gethostbyname` zurückgegeben werden, und nicht nur die erste IP-Adresse.

Wenn Sie in Ihrer Anwendung an allen Stellen konsistente IP-Adressen benötigen, konfigurieren Sie Anwendung so, dass die Pro-Knoten-Adresse sowohl an die Client- als auch an die Serverseite angebunden ist; so scheinen alle Verbindungen über die Pro-Knoten-Adresse zu laufen.

Ressourcen, Ressourcengruppen und Ressourcentypen

Datendienste verwenden mehrere Typen von *Ressourcen*: Anwendungen wie Apache Web Server oder Sun ONE Web Server verwenden Netzwerkadressen (logische Hostnamen und gemeinsam genutzte Adressen), von denen die Anwendungen abhängen. Anwendung und Netzwerkressourcen bilden eine grundlegende Einheit, die vom RGM verwaltet wird.

Datendienste sind Ressourcentypen. Sun Cluster HA für Oracle ist zum Beispiel der Ressourcentyp `SUNW.oracle-server` und Sun Cluster HA für Apache der Ressourcentyp `SUNW.apache`.

Eine Ressource ist eine Instanz eines *Ressourcentyps*, der Cluster-weit definiert ist. Es gibt mehrere definierte Ressourcentypen.

Netzwerkressourcen sind entweder `SUNW.LogicalHostname`- oder `SUNW.SharedAddress`-Ressourcentypen. Diese beiden Ressourcentypen sind in der Sun Cluster-Software vorregistriert.

Die Ressourcentypen `SUNW.HAStorage` und `HAStoragePlus` werden zur Synchronisierung beim Starten von Ressourcen und Plattengerätegruppen verwendet, von denen die Ressourcen abhängen. Damit wird sichergestellt, dass die Pfade zu Einhängepunkten im Cluster-Dateisystem, globalen Geräten und Gerätegruppennamen verfügbar sind, bevor ein Datendienst startet. Weitere Informationen fnden Sie unter "Synchronizing the Startups Between Resource Groups and Disk Device Group" im *Data Services Installation and Configuration Guide*. (Der `HAStoragePlus`-Ressourcentyp ist ab Sun Cluster 3.0 5/02 verfügbar und damit eine weitere Funktion, um lokale Dateisysteme hoch verfügbar zu machen. Weitere Informationen zu dieser Funktion finden Sie unter „`HAStoragePlus`-Ressourcentyp" auf Seite 49.)

RGM-verwaltete Ressourcen werden in Gruppen zusammengefasst, die als *Ressourcengruppen* bezeichnet werden, so dass sie als Einheit verwaltet werden können. Eine Ressourcengruppe migriert als Einheit, wenn ein Failover oder ein Switchover für die Ressourcengruppe eingeleitet wird.

Hinweis – Wenn Sie eine Ressourcengruppe mit Anwendungsressourcen online bringen, wird die Anwendung gestartet. Die Datendienst-Startmethode wartet, bis die Anwendung aktiv ist und ausgeführt wird, bevor sie erfolgreich beendet wird. Die Feststellung, wann die Anwendung aktiv ist und ausgeführt wird, erfolgt auf dieselbe Weise, in der Datendienst-Fehler-Monitor feststellt, ob ein Datendienst Clients bedient. Weitere Informationen zu diesem Prozess finden Sie im *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Ressourcengruppen-Manager (RGM)

Der RGM steuert Datendienste (Anwendungen) als Ressourcen, die mit *Ressourcentypen*-Implementierungen verwaltet werden. Diese Implementierungen werden entweder von Sun geliefert oder von einem Entwickler mithilfe einer generischen Datendienstvorlage, der API für die Datendienst-Entwicklungsbibliothek (DSDL-API) oder der Ressourcenverwaltungs-API (RMAPI), erstellt. Der

Cluster-Verwalter erstellt und verwaltet die Ressourcen in Containern, die als *Ressourcengruppen* bezeichnet werden. Der RGM stoppt und startet die Ressourcengruppen auf den ausgewählten Knoten je nach den Änderungen in der Cluster-Mitgliedschaft.

Der RGM verwaltet *Ressourcen* und *Ressourcengruppen*. RGM-Aktionen ändern den Zustand von Ressourcen und Ressourcengruppen von online zu offline und umgekehrt. Eine umfassende Beschreibung dieser Zustände und der Einstellungen für Ressourcen und Ressourcengruppen finden Sie im Abschnitt „Zustände und Einstellungen für Ressourcen und Ressourcengruppen“ auf Seite 74. Informationen über das Starten des Ressourcenverwaltungsprojekts unter der Steuerung des RGM finden Sie unter „Ressourcen, Ressourcengruppen und Ressourcentypen“ auf Seite 72.

Zustände und Einstellungen für Ressourcen und Ressourcengruppen

Ein Verwalter verwendet statische Einstellungen für Ressourcen und Ressourcengruppen. Diese Einstellungen können nur über Verwaltungsaktionen geändert werden. Der RGM ändert die dynamischen “Zustände” von Ressourcengruppen. Diese Einstellungen und Zustände sind in der nachstehenden Liste beschrieben.

- Verwaltet oder unverwaltet – Diese Cluster-weiten Einstellungen gelten nur für Ressourcengruppen. Ressourcengruppen werden vom RGM verwaltet. Mit dem `scrgadm(1M)`-Befehl wird die Verwaltung einer Ressourcengruppe durch den RGM aktiviert bzw. deaktiviert. Diese Einstellungen werden bei einer Cluster-Neukonfiguration nicht geändert.

  Wenn eine Ressourcengruppe zum ersten Mal erstellt wird, ist sie unverwaltet. Sie muss verwaltet werden, bevor enthaltene Ressourcen aktiv werden können.

  Bei einigen Datendiensten, zum Beispiel einem Scalable-Webserver, muss diese Arbeit vor dem Starten und nach dem Anhalten der Netzwerkressourcen erledigt werden. Dies wird durch die Datendienstmethoden zum Initialisieren (INIT) und Beenden (FINI) erledigt. Die INIT-Methoden werden nur ausgeführt, wenn die Ressourcengruppe, zu der die Ressource gehört, den Zustand "Verwaltet" hat.

  Wenn der Status einer Ressourcengruppe von unverwaltet auf verwaltet gesetzt wird, werden alle registrierten INIT-Methoden für die in der Gruppe enthaltenen Ressourcen ausgeführt.

  Wenn der Status einer Ressourcengruppe von verwaltet auf unverwaltet gesetzt wird, werden alle registrierten FINI-Methoden aufgerufen, um eine Bereinigung durchzuführen.

  Am häufigsten werden INIT- und FINI-Methoden für Netzwerkressourcen für Scalable-Dienste eingesetzt, aber sie können für alle beliebigen Initialisierungs- oder Bereinigungsaufgaben eingesetzt werden, die nicht von der Anwendung erledigt werden.

- Aktiviert oder deaktiviert – Diese Cluster-weiten Einstellungen gelten für Ressourcen. Der `scrgadm(1M)`-Befehl kann zur Aktivierung oder Deaktivierung einer Ressource verwendet werden. Diese Einstellungen werden bei einer Cluster-Neukonfiguration nicht geändert.

  Bei einer normalen Einstellung ist eine Ressource aktiviert und wird aktiv im System ausgeführt.

  Wenn die Ressource aus bestimmten Gründen auf keinem Cluster-Knoten verfügbar sein soll, deaktivieren Sie die Ressource. Eine deaktivierte Ressource steht nicht zur allgemeinen Verwendung zur Verfügung.

- Online oder offline – Diese dynamischen Zustände gelten sowohl für Ressourcen als auch für Ressourcengruppen.

  Diese Zustände ändern sich bei den Cluster-Übergangen im Verlauf der Cluster-Rekonfigurationsschritte bei einem Switchover oder Failover. Sie können auch durch Verwaltungsaktionen geändert werden. Mit `scswitch(1M)` können Sie den Zustand einer Ressource oder einer Ressourcengruppe von online zu offline und umgekehrt ändern.

  Eine Failover-Ressource oder -Ressourcengruppe kann nur auf jeweils einem Knoten gleichzeitig online sein. Eine Scalable-Ressource oder -Ressourcengruppe kann gleichzeitig auf mehreren Knoten online und auf anderen offline sein. Während eines Switchovers oder Failovers werden Ressourcengruppen mit den darin enthaltenen Ressourcen auf einem Knoten offline genommen und auf einem anderen Knoten online gebracht.

  Wenn eine Ressourcengruppe offline ist, sind alle dazugehörigen Ressourcen offline. Wenn eine Ressourcengruppe online ist, sind alle dazugehörigen aktivierten Ressourcen online.

  Ressourcengruppen können mehrere Ressourcen mit Abhängigkeiten zwischen Ressourcen enthalten. Aufgrund dieser Abhängigkeiten müssen die Ressourcen in einer bestimmten Reihenfolge online gebracht und offline genommen werden. Die Methoden, um die Ressourcen online zu bringen und offline zu nehmen, können für jede Ressource unterschiedlich lange dauern. Die Zustände online und offline können für Ressourcen einer einzigen Ressourcengruppe aufgrund von Ressourcenabhängigkeiten und Unterschieden bei Start- und Stoppzeit bei einer Cluster-Rekonfiguration unterschiedlich sein.

Ressourcen- und Ressourcengruppeneigenschaften

Sie können Eigenschaftenwerte für die Ressourcen und Ressourcengruppen Ihrer SunPlex-Datendienste konfigurieren. Standardeigenschaften sind für alle Datendienste gleich. Erweiterungseigenschaften sind für jeden Datendienst spezifisch. Einige Standard- und Erweiterungseigenschaften sind mit Standardeinstellungen konfiguriert, so dass Sie diese nicht ändern müssen. Andere müssen im Rahmen der Erstellung und Konfiguration von Ressourcen eingestellt werden. Die Dokumentation für jeden Datendienst gibt an, welche Ressourceneigenschaften eingestellt werden können und wie diese einzustellen sind.

Die Standardeigenschaften werden zur Konfiguration von Ressourcen- und Ressourcengruppeneigenschaften eingesetzt, die in der Regel von keinem bestimmten Datendienst abhängen. Die Standardeigenschaften werden in einem Anhang im *Sun Cluster 3.1 Data Services Installation and Configuration Guide* beschrieben.

Die RGM-Erweiterungseigenschaften liefern Informationen wie zum Beispiel den Speicherort von binären Anwendungsdateien und Konfigurationsdateien. Sie ändern die Erweiterungseigenschaften beim Konfigurieren der Datendienste. Die Erweiterungseigenschaften werden im Kapitel zum Datendienst im *Sun Cluster 3.1 Data Services Installation and Configuration Guide* beschrieben.

Datendienst-Projektkonfiguration

Datendienste können für einen Start mit einem Solaris-Projektnamen konfiguriert werden, wenn sie mit dem RGM online gebracht werden. Die Konfiguration ordnet einer vom RGM verwalteten Ressource oder Ressourcengruppe eine Solaris Projekt-ID zu. Durch die Zuordnung der Ressource oder Ressourcengruppe zu einer Projekt-ID können Sie anspruchsvolle Steuerfunktionen, die in der Solaris-Umgebung verfügbar sind, für die Verwaltung von Arbeitslasten und Verbrauch in Ihrem Cluster einsetzen.

Hinweis – Diese Konfiguration können Sie nur dann durchführen, wenn Sie mit der aktuellen Version der Sun Cluster-Software unter Solaris 9 arbeiten.

Mithilfe der Solaris-Verwaltungsfunktion in einer Cluster-Umgebung können Sie sicherstellen, dass die wichtigsten Anwendungen Priorität erhalten, wenn sie einen Knoten mit anderen Anwendungen teilen. Anwendungen können einen Knoten teilen, wenn Sie mit konsolidierten Diensten arbeiten oder weil Anwendungen bei einem Failover den Knoten gewechselt haben. Die Verwendung der hier beschriebenen Verwaltungsfunktion kann die Verfügbarkeit einer kritischen Awendung verbessern, indem andere Anwendungen mit niedrigerer Priorität daran gehindert werden, zu viele Systemleistungen wie CPU-Zeit in Anspruch zu nehmen.

Hinweis – In der Solaris-Dokumentation zu dieser Funktion werden CPU-Zeit, Prozesse, Aufgaben und vergleichbare Komponenten als 'Ressourcen' beschrieben. In der Sun Cluster-Dokumentation hingegen wird der Begriff 'Ressourcen' zur Beschreibung von Entitäten verwendet, die vom RGM gesteuert werden. Im nachstehenden Abschnitt wird der Begriff 'Ressource' für Sun Cluster-Entitäten verwendet, die vom RGM gesteuert werden, und der Begriff 'Leistungen' für CPU-Zeit, Prozesse und Aufgaben.

Dieser Abschnitt enthält eine konzeptionelle Beschreibung zur Konfiguration von Datendiensten, um die Prozesse in einem gegebenen Solaris 9 `project`(4) zu starten. In diesem Abschnitt werden auch mehrere Failover-Szenarien beschrieben. Des

Weiteren enthält er Empfehlungen zur Planung, wie die in der Solaris-Umgebung zur Verfügung gestellte Verwaltungsfunktion eingesetzt werden kann. Detaillierte konzeptionelle und verfahrenstechnische Dokumentation zur Verwaltungsfunktion finden Sie im *System Administration Guide: Resource Management and Network Services* aus der *Solaris 9 System Administrator Collection.*

Bei der Konfiguration von Ressourcen und Ressourcengruppen zur Verwendung der Solaris-Verwaltungsfunktion in einem Cluster können Sie folgenden Prozess auf höchster Ebene einsetzen:

1. Konfigurieren von Anwendungen als Teil der Ressource.
2. Konfigurieren von Ressourcen als Teil einer Ressourcengruppe.
3. Aktivieren von Ressourcen in der Ressourcengruppe.
4. Aktivieren der Verwaltung einer Ressourcengruppe.
5. Erstellen eines Solaris-Projekts für die Ressourcengruppe.
6. Konfigurieren von Standardeigenschaften, um den Ressourcengruppennamen dem unter Punkt 5 erstellten Projekt zuzuordnen.
7. Online-bringen der Ressourcengruppe.

Verwenden Sie zur Konfiguration der Standardeigenschaften `Resource_project_name` oder `RG_project_name` die `-y`-Option mit dem `scrgadm`(1M)-Befehl, um die Solaris-Projekt-ID der Ressource oder Ressourcengruppe zuzuordnen. Stellen Sie die Eigenschaftenwerte für die Ressource oder die Ressourcengruppe ein. Die Definitionen der Eigenschaften finden Sie unter “Standard Properties” in *Sun Cluster 3.1 Data Services Installation and Configuration Guide*. Siehe auch `r_properties`(5) und `rg_properties`(5) for property descriptions.

Der angegebene Projektname muss in der Projektdatenbank (`/etc/Projekt`) vorhanden sein, und der Root-Benutzer muss als Mitglied des genannten Projekts konfiguriert sein. Konzeptionelle Informationen zur Projektnamen-Datenbank finden Sie unter “Projects and Tasks” in *System Administration Guide: Resource Management and Network Services* in der *Solaris 9 System Administrator Collection*. Eine Beschreibung der Syntax der Projektdatei finden Sie unter `project`(4).

Wenn der RGM Ressourcen oder Ressourcengruppen online bringt, startet er die verknüpften Prozesse unter dem Projektnamen.

Hinweis – Benutzer können die Ressource oder Ressourcengruppe jederzeit einem Projekt zuordnen. Der neue Projektname ist jedoch erst gültig, wenn die Ressource oder Ressourcengruppe mit dem RGM offline genommen und wieder online gebracht wird.

Durch das Starten von Ressourcen und Ressourcengruppen unter dem Projektnamen können Sie die folgenden Funktionen zur Verwaltung der Systemleistungen in Ihrem Cluster konfigurieren.

- Erweiterte Berechnung – Ein flexibler Weg zur Aufzeichnung des Verbrauchs auf Aufgaben- oder Prozessbasis. Mit dieser Funktion können Sie die historische Nutzung nachverfolgen und den Leistungsbedarf für zukünftige Arbeitslasten beurteilen.
- Steuerungen – Ein Mechanismus zur Einschränkung von Systemleistungen. Damit kann verhindert werden, dass Prozesse, Aufgaben und Projekte große Anteile von bestimmten Systemleistungen verbrauchen.
- Faire Nutzungsplanung (FSS) – Damit besteht die Möglichkeit, die Zuweisung der verfügbaren CPU-Zeit auf die Arbeitslasten je nach ihrer Bedeutung zu steuern. Die Bedeutung der Arbeitslast wird durch die Anteile an CPU-Zeit ausgedrückt, die Sie jeder Arbeitslast zuweisen. Eine Beschreibung der Befehlszeile, um FSS als Ihren Standardplaner festzulegen, finden Sie unter `dispadmin`(1M). Weitere Information finden Sie auch unter `priocntl`(1), `ps`(1) und `FSS`(7).
- Pools – Damit besteht die Möglichkeit, Partitionen je nach Anwendungsanforderungen für interaktive Anwendungen zu nutzen. Pools können zur Partitionierung eines Servers eingesetzt werden, der eine Reihe unterschiedlicher Softwareanwendungen unterstützt. Die Verwendung von Pools führt zu einer besser vorhersehbaren Antwort für jede Anwendung.

Bestimmen der Anforderungen der Projektkonfiguration

Bevor Sie die Datendienste zur Verwendung der von Solaris zur Verfügung gestellten Steuermöglichkeiten in einer Sun Cluster-Umgebung konfigurieren, müssen Sie entscheiden, wie Sie die Ressourcen bei Switchover- oder Failover-Vorgängen steuern und verfolgen möchten. Identifizieren Sie vor der Konfiguration eines neuen Projekts die Abhängigkeiten innerhalb Ihres Clusters. Ressourcen und Ressourcengruppen hängen zum Beispiel von Plattengerätegruppen ab. Verwenden Sie die Ressourcengruppeneigenschaften `nodelist`, `failback`, `maximum_primaries` und `desired_primaries`, die mit `scrgadm`(1M) konfiguriert werden, um die Prioritäten in der Knotenliste für Ihre Ressourcengruppe zu identifizieren. Eine kurze Erläuterung der Abhängigkeiten zwischen Ressourcengruppen und Plattengerätegruppen finden Sie unter “Relationship Between Resource Groups and Disk Device Groups” in *Sun Cluster 3.1 Data Services Installation and Configuration Guide*. Detaillierte Beschreibungen der Eigenschaften finden Sie unter `rg_properties`(5).

Legen Sie die Prioritäten der Knotenliste für die Plattengerätegruppe mit den Eigenschaften `preferenced` und `failback` fest, die mit `scrgadm`(1M) und `scsetup`(1M) konfiguriert werden. Verfahrenstechnische Informationen finden Sie unter “So ändern Sie Plattengeräteeigenschaften” unter “ Administering Disk Device Groups” in *Sun Cluster 3.1 10/03 Handbuch Systemverwaltung*. Konzeptionelle Informationen zur Knotenkonfiguration und dem Verhalten von Failover- und Scalable-Datendiensten finden Sie unter „SunPlex-System-Hardwarekomponenten“ auf Seite 21.

Wenn Sie alle Cluster-Knoten gleich konfigurieren, werden die Nutzungsgrenzen auf Primär- und Sekundärknoten identisch durchgesetzt. Die Konfigurationsparameter der Projekte müssen nicht für alle Anwendungen in den Konfigurationsdateien auf

allen Knoten gleich sein. Auf alle der Anwendung zugeordneten Projekte muss zumindest die Projektdatenbank auf allen potenziellen Mastern der Anwendung zugreifen können. Angenommen, der Master für Anwendung 1 ist *phys-schost-1*, kann aber durch ein Switchover oder Failover auch *phys-schost-2* oder *phys-schost-3* sein. Auf allen drei Knoten (*phys-schost-1*, *phys-schost-2* und *phys-schost-3*) muss Zugriff auf das der Anwendung 1 zugeordnete Projekt bestehen.

Hinweis – Die Projektdatenbank-Informationen können eine lokale `/etc/project`-Datenbankdatei sein oder in der NIS-Karte oder im LDAP-Verzeichnisdienst gespeichert sein.

Die Solaris-Umgebung ermöglicht eine flexible Konfiguration der Nutzungsparameter, und Sun Cluster erlegt nur wenige Einschränkungen auf. Die Konfigurationsauswahl hängt von den Standortbedürfnissen ab. Beachten Sie die allgemeinen Richtlinien in den nachfolgenden Abschnitten, bevor Sie Ihre Systeme konfigurieren.

Einstellen virtueller Speicherbegrenzungen für Prozesse

Stellen Sie die `process.max-address-space`-Steuerung ein, um den virtuellen Speicher auf Prozessbasis zu begrenzen. Detaillierte Informationen zur Einstellung des `process.max-address-space`-Wertes finden Sie unter `rctladm`(1M).

Konfigurieren Sie bei Verwendung von Verwaltungssteuerungen die Speicherbegrenzung angemessen, um unnötige Failover-Vorgänge von Anwendungen und einen "Ping-Pong"-Effekt bei den Anwendungen zu vermeiden. Im Allgemeinen gilt Folgendes:

- Stellen Sie die Speicherbegrenzungen nicht zu niedrig ein.

  Wenn eine Anwendung die Speicherbegrenzung erreicht, kann ein Failover erfolgen. Diese Richtlinie ist besonders für Datenbankanwendungen von Bedeutung, wo das Erreichen einer virtuellen Speicherbegrenzung unerwartete Folgen haben kann.

- Stellen Sie die Speicherbegrenzungen für Primär- und Sekundärknoten nicht identisch ein.

  Identische Begrenzungen können einen Ping-Pong-Effekt auslösen, wenn eine Anwendung die Speichergrenze erreicht und ein Failover auf einen Sekundärknoten mit einer identischen Speicherbegrenzung erfolgt. Stellen Sie die Speicherbegrenzung auf dem Sekundärknoten etwas höher ein. Network MultipathingDer Unterschied bei den Speicherbegrenzungen trägt zur Vermeidung des Ping-Pong-Szenarios bei und gibt dem Systemverwalter Zeit, die Parameter nach Bedarf zu korrigieren.

- Verwenden Sie die Speicherbegrenzung der Ressourcenverwaltung für den Lastausgleich.

Sie können zum Beispiel eine fehlerhaft arbeitende Anwendung mit Speicherbegrenzungen davon abhalten, zu viel Auslagerungsspeicher zu belegen.

Failover-Szenarien

Sie können die Verwaltungsparameter so konfigurieren, dass die Zuweisung in der Projektkonfiguration (`/etc/project`) im normalen Cluster-Betrieb und in Switchover- und Failover-Situationen funktioniert.

Die nachfolgenden Abschnitte sind als Beispiel gedachte Szenarien.

- In den ersten beiden Abschnitte, "Zwei-Knoten-Cluster mit zwei Anwendungen" und "Zwei-Knoten-Cluster mit drei Anwendungen," werden Failover-Szenarien für ganze Knoten beschrieben.
- Der Abschnitt "Failover der Ressourcengruppe" illustriert den Failover-Vorgang einer einzigen Anwendung.

In einer Cluster-Umgebung wird eine Anwendung als Teil einer Ressource und eine Ressource als Teil einer Ressourcengruppe (RG) konfiguriert. Wenn ein Fehler auftritt, wechselt die Ressourcengruppe zusammen mit den zugeordneten Anwendungen auf einen anderen Knoten. In den nachstehenden Beispielen sind die Ressourcen nicht explizit angegeben. Angenommen, jede Ressource umfasst nur eine Anwendung.

Hinweis – Ein Failover erfolgt in der bevorzugten Reihenfolge, wie sie in der Knotenliste im RGM angegeben ist.

Für die nachstehenden Beispiele gelten folgende Beschränkungen:

- Anwendung 1 (Anw-1) ist in der Ressourcengruppe RG-1 konfiguriert.
- Anwendung 2 (Anw-2) ist in der Ressourcengruppe RG-2 konfiguriert.
- Anwendung 3 (Anw-3) ist in der Ressourcengruppe RG-3 konfiguriert.

Obwohl die Anzahl der zugeordneten Anteile gleich bleibt, ändert sich der jeder Anwendung zugewiesene Prozentsatz an CPU-Zeit nach dem Failover. Dieser Prozentsatz hängt von der Anzahl der auf dem Knoten ausgeführten Anwendungen sowie von der Anzahl an Anteilen ab, die jeder aktiven Anwendung zugeordnet sind.

In diesen Szenarien werden folgende Konfigurationen vorausgesetzt.

- Alle Anwendungen sind unter einem gemeinsamen Projekt konfiguriert.
- Jede Ressource enthält nur eine einzige Anwendung.
- Die Anwendungen sind die einzigen aktiven Prozesse auf den Knoten.
- Die Projektdatenbanken sind auf allen Knoten des Clusters gleich konfiguriert.

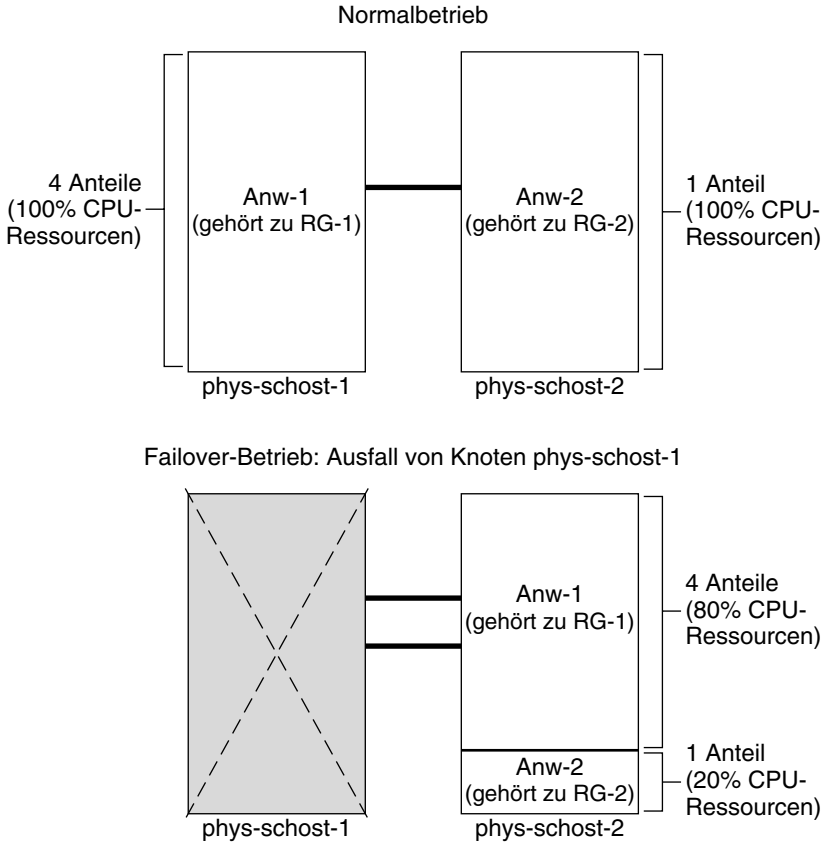
Zwei-Knoten-Cluster mit zwei Anwendungen

Auf einem Zwei-Cluster-Knoten können Sie zwei Anwendungen konfigurieren, um sicherzustellen, dass jeder reale Host (*phys-schost-1*, *phys-schost-2*) als Standardmaster für eine Anwendung fungiert. Jeder reale Host fungiert als Sekundärknoten für den anderen realen Host. Alle Projekte, die Anwendung 1 und Anwendung 2 zugeordnet sind, müssen in den Datenbankdateien der Projekte auf beiden Knoten vertreten sein. Im normalen Cluster-Betrieb wird jede Anwendung auf dem eigenen Standard-Master ausgeführt und erhält über die Verwaltungsfunkton die gesamte CPU-Zeit zugewiesen.

Nach einem Failover oder Switchover werden beide Anwendungen auf einem einzigen Knoten ausgeführt und erhalten die Anteile zugewiesen, die in der Konfigurationsdatei angegeben wurden. Dieser Eintrag in der `/etc/project`-Datei gibt zum Beispiel an, dass Anwendung 1 vier Anteile und Anwendung 2 ein Anteil zugewiesen wird.

```
Prj_1:100:project for App-1:root::project.cpu-shares=(privileged,4,none)
Prj_2:101:project for App-2:root::project.cpu-shares=(privileged,1,none)
```

Die nachstehende Abbildung illustriert den normalen und den Failover-Betrieb in dieser Konfiguration. Die Anzahl der zugewiesenen Anteile ändert sich nicht. Der Prozentsatz an CPU-Zeit, der jeder Anwendung zur Verfügung steht, kann sich jedoch je nach Anzahl der Anteile ändern, die jedem Prozess mit CPU-Zeit-Bedarf zugewiesen werden.



Zwei-Knoten-Cluster mit drei Anwendungen

Auf einem Zwei-Knoten-Cluster mit drei Anwendungen können Sie einen realen Host (*phys-schost-1*) als Standard-Master für eine Anwendung und den zweiten realen Host (*phys-schost-2*) als Standard-Master für die restlichen zwei Anwendungen konfigurieren. Im folgenden Beispiel wird davon ausgegangen, dass die Projekt-Datenbankdatei auf jedem Knoten vorhanden ist. Die Projekt-Datenbankdatei wird bei einem Failover oder Switchover nicht geändert.

```
Prj_1:103:project for App-1:root::project.cpu-shares=(privileged,5,none)
Prj_2:104:project for App_2:root::project.cpu-shares=(privileged,3,none)
Prj_3:105:project for App_3:root::project.cpu-shares=(privileged,2,none)
```

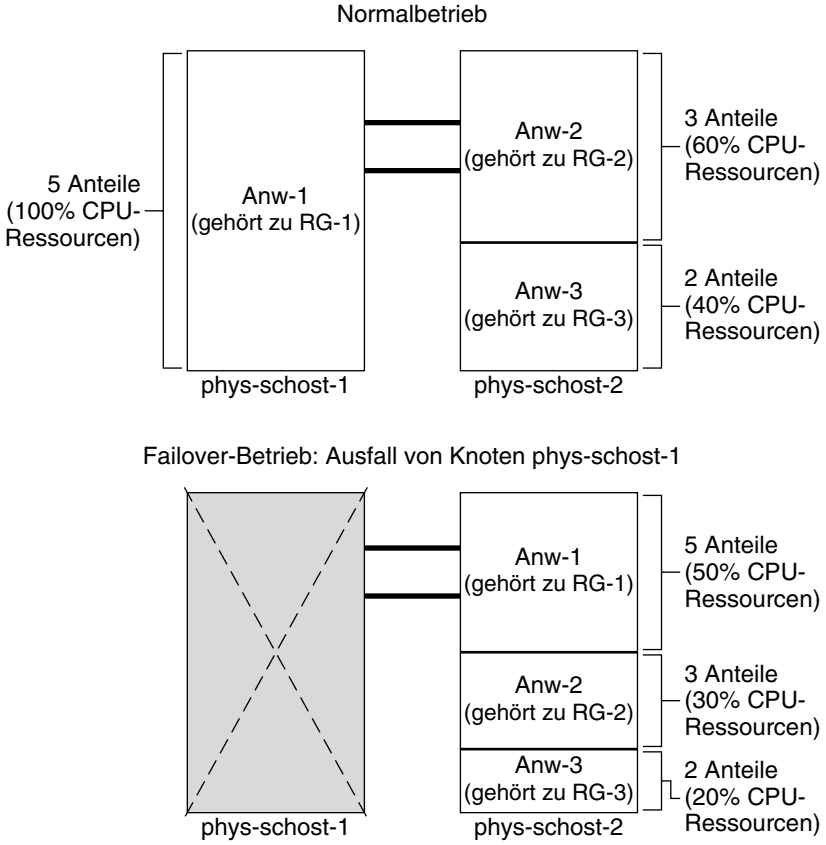
Im normalen Cluster-Betrieb sind der Anwendung 1 fünf Anteile auf dem entsprechenden Standard-Master *phys-schost-1* zugewiesen. Dies entspricht 100 Prozent der CPU-Zeit, weil es die einzige Anwendung ist, die CPU-Zeit auf diesem

Knoten benötigt. Den Anwendungen 2 und 3 sind jeweils drei und zwei Anteile auf dem dazugehörigen Standard-Master *phys-schost-2* zugewiesen. Damit entfällt auf Anwendung 2 60 Prozent der CPU-Zeit und auf Anwendung 3 40 Prozent der CPU-Zeit im normalen Betrieb.

Wenn Anwendung 1 bei einem Failover oder Switchover auf *phys-schost-2* wechselt, bleiben die Anteile für alle drei Anwendungen gleich. Aber der Prozentsatz der CPU-Ressourcen wird anhand der Projekt-Datenbankdatei neu zugewiesen.

- Anwendung 1 erhält mit 5 Anteilen 50 Prozent der CPU.
- Anwendung 2 erhält mit 3 Anteilen 30 Prozent der CPU.
- Anwendung 3 erhält mit 2 Anteilen 20 Prozent der CPU.

Die nachstehende Abbildung illustriert den normalen und den Failover-Betrieb in dieser Konfiguration.



Failover der Ressourcengruppe

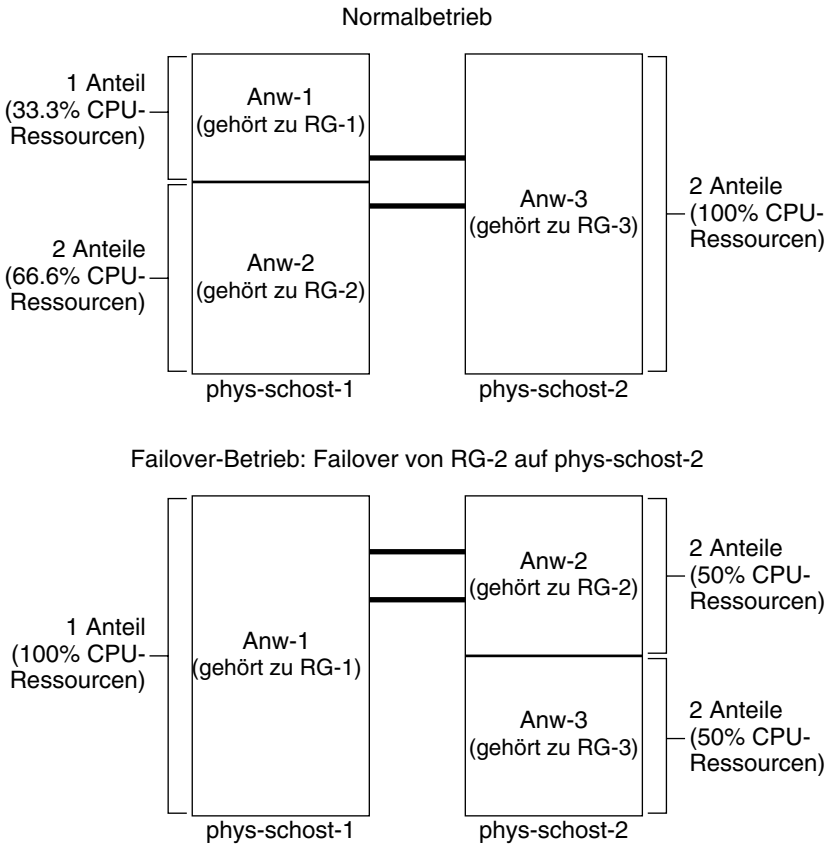
In einer Konfiguration, in der mehrere Ressourcengruppen denselben Standard-Master haben, kann eine Ressourcengruppe (und die zugeordneten Anwendungen) bei einem Failover oder Switchover auf einen Sekundärknoten wechseln. Dabei läuft der Standard-Master im Cluster weiter.

Hinweis – Die Anwendung, die den Failover-Vorgang durchführt, erhält dabei die Ressourcen auf dem Sekundärknoten zugewiesen, die in der Konfigurationsdatei angegeben sind. In diesem Beispiel haben die Projekt-Datenbankdateien auf Primär- und Sekundärknoten dieselbe Konfiguration.

Diese Beispielskonfigurationsdatei gibt an, dass Anwendung 1 ein Anteil, Anwendung 2 zwei Anteile und Anwendung 3 zwei Anteile zugewiesen werden.

```
Prj_1:106:project for App_1:root::project.cpu-shares=(privileged,1,none)
Prj_2:107:project for App_2:root::project.cpu-shares=(privileged,2,none)
Prj_3:108:project for App_3:root::project.cpu-shares=(privileged,2,none)
```

Die nachstehende Abbildung illustriert den normalen und den Failover-Betrieb in dieser Konfiguration, wobei RG-2 mit Anwendung 2 zu *phys-schost-2* wechselt. Beachten Sie, dass sich die Anzahl der zugewiesenen Anteile nicht ändert. Der Prozentsatz an CPU-Zeit, der jeder Anwendung zur Verfügung steht, kann sich jedoch je nach Anzahl der Anteile ändern, die jeder Anwendung mit CPU-Zeit-Bedarf zugewiesen werden.



Öffentliche Netzwerkadapter und IP Network Multipathing

Clients führen über das öffentliche Netzwerk Datenanforderungen an den Cluster durch. Jeder Cluster-Knoten ist über ein Paar öffentlicher Netzwerkadapter mit mindestens einem öffentlichen Netzwerk verbunden.

Die Solaris IPMP-Software (Internet Protocol Network Multipathing) von Sun Cluster liefert den Basismechanismus für die Überwachung öffentlicher Netzwerkadapter und das Wechseln von IP-Adressen von einem Adapter auf einen anderen, wenn ein Fehler erkannt wird. Jeder Cluster-Knoten hat seine eigene IP Network Multipathing-Konfiguration, die sich von der auf anderen Cluster-Knoten unterscheiden kann.

Öffentliche Netzwerkadapter sind als *IPMP-Gruppen* (Multipathing-Gruppen) organisiert. Jede Multipathing-Gruppe hat mindestens einen öffentlichen Netzwerkadapter. Jeder Adapter in einer Multipathing-Gruppe kann aktiv sein oder Sie können Standby-Schnittstellen konfigurieren, die bis zum Auftreten eines Failovers inaktiv bleiben. Der `in.mpathd`-Mutlipathing-Dämon verwendet eine IP-Testadresse, um Fehler und Reparaturen zu erkennen. Erkennt der Multipathing-Dämon einen Fehler auf einem der Adapter, erfolgt ein Failover. Der gesamte Netzwerkzugriff geht vom fehlerhaften Adapter auf einen anderen funktionsfähigen Adapter in der Multipathing-Gruppe unter Aufrechterhaltung der Konnektivität des öffentlichen Netzwerkes für den Knoten über. Wenn eine Standby-Schnittstelle konfiguriert wurde, wählt der Dämon die Standby-Schnittstelle aus. Andernfalls wählt `in.mpathd` die Schnittstelle mit der niedrigsten Anzahl von IP-Adressen aus. Da das Failover auf Adapterschnittstellenebene erfolgt, sind Verbindungen auf höherer Ebene wie TCP mit Ausnahme einer kurzen Zeitspanne während des Failover-Vorgangs nicht betroffen. Wenn das Failover der IP-Adressen erfolgreich abgeschlossen ist, werden freie ARP-Übertragungen gesendet. Die Konnektivität der Remote-Clients bleibt damit erhalten.

Hinweis – Aufgrund der TCP-Wiederherstellungseigenschaften bei Stau kann an TCP-Endpunkten nach einem erfolgreichen Failover eine weitere Verzögerung entstehen, da einige Segmente beim Failover verloren gehen können und der Stausteuermechanismus im TCP aktiviert wird.

Multipathing-Gruppen bilden die Bausteine für logische Hostnamen und gemeinsam genutzte Adressen. Sie können auch unabhängig von logischen Hostnamen und gemeinsam genutzten Adressen Multipathing-Gruppen erstellen, um die Konnektivität der Cluster-Knoten mit dem öffentlichen Netzwerk zu überwachen. Dieselbe Multipathing-Gruppe kann auf einem Knoten eine beliebige Anzahl logischer Hostnamen oder gemeinsam genutzter Adressen hosten. Weitere Information zu den Ressourcen logischer Hostname und gemeinsam genutzte Adressen finden Sie im *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Hinweis – Der IP Network Multipathing-Mechanismus ist auf das Erkennen und Verbergen von Adapterfehlern ausgelegt. Zweck dieser Gestaltung ist nicht, die Wiederherstellung von einem Verwalter durchzuführen und mit `ifconfig(1M)` eine der logischen (oder gemeinsam genutzten) IP-Adressen zu entfernen. Die Sun Cluster-Software zeigt die logischen und gemeinsam genutzten IP-Adressen als Ressourcen an, die vom RGM verwaltet werden. Ein Verwalter muss eine IP-Adresse mit `scrgadm`(1M) entfernen oder hinzufügen, um die Ressourcengruppe mit der Ressource zu ändern.

Weitere Informationen zur Solaris-Implementierung von IP Network Multipathing finden Sie in der entsprechenden Dokumentation zur Solaris-Betriebsumgebung, die auf Ihrem Cluster installiert ist.

Version des Betriebssystems	Anweisungen siehe
Solaris 8-Betriebsumgebung	*IP Network Multipathing Administration Guide*
Solaris 9-Betriebsumgebung	"IP Network Multipathing Topics " in *System Administration Guide: IP Services*

Unterstützung der dynamischen Rekonfiguration

Die Softwarefunktion zur Unterstützung der dynamischen Rekonfiguration (DR) in Sun Cluster 3.1 wird schrittweise umgesetzt. In diesem Abschnitt werden Konzepte und Erwägungen im Hinblick auf die Unterstützung der DR-Funktion durch Sun Cluster 3.1 beschrieben.

Beachten Sie, dass alle für die DR-Funktion von Solaris dokumentierten Anforderungen, Verfahren und Einschränkungen auch für die DR-Unterstützung von Sun Cluster gelten (mit Ausnahme des Vorgangs zur Stilllegung der Betriebsumgebung). Sehen Sie deswegen die Dokumentation zur Solaris DR-Funktion nochmals durch, *bevor* Sie die DR-Funktion mit der Sun Cluster-Software verwenden. Lesen Sie insbesondere nochmals die Themen, die sich mit nicht vernetzten E/A-Geräten während eines DR-Trennungsvorgangs beschäftigen. Das *Sun Enterprise 10000 Dynamic Reconfiguration User Guide* und das *Sun Enterprise 10000 Dynamic Reconfiguration Reference Manual* (aus den Dokumentationsreihen *Solaris 8 on Sun Hardware* oder *Solaris 9 on Sun Hardware*) können von http://docs.sun.com heruntergeladen werden.

Allgemeine Beschreibung der dynamischen Rekonfiguration

Die DR-Funktion ermöglicht bestimmte Vorgänge wie zum Beispiel das Entfernen von Systemhardware bei laufenden Systemen. Die DR-Prozesse sind so ausgelegt, dass sie einen kontinuierlichen Systembetrieb sicherstellen, ohne das System anhalten oder die Cluster-Verfügbarkeit unterbrechen zu müssen.

DR arbeitet auf Board-Ebene. Deswegen hat ein DR-Vorgang Auswirkungen auf alle Komponenten eines Boards. Jedes Board kann mehrere Komponenten enthalten, einschließlich CPUs, Speicher und Peripherieschnittstellen für Plattenlaufwerke, Bandlaufwerke und Netzwerkverbindungen.

Das Entfernen eines Boards mit aktiven Komponenten führt zu Systemfehlern. Vor dem Entfernen eines Boards fragt das DR-Teilsystem andere Teilsysteme wie Sun Cluster ab, um festzustellen, ob die Komponenten des Boards genutzt werden. Wenn das DR-Teilsystem eine Board-Nutzung feststellt, wird der DR-Vorgang zur Board-Entfernung nicht ausgeführt. Ein DR-Vorgang zur Board-Entfernung ist insofern immer sicher, als das DR-Teilsystem Vorgänge an Boards mit aktiven Komponenten ablehnt.

Der DR-Vorgang zur Board-Hinzufügung ist ebenfalls immer sicher. CPUs und Speicher auf einem neu hinzugefügten Board werden automatisch vom System verfügbar gemacht. Der Systemverwalter muss den Cluster jedoch manuell konfigurieren, um Komponenten des neu hinzugefügten Boards aktiv nutzen zu können.

Hinweis – Das DR-Teilsystem hat mehrere Ebenen. Wenn ein Fehler auf einer unteren Ebene gemeldet wird, meldet auch die übergeordnete Ebene einen Fehler. Die untere Ebene meldet den spezifischen Fehler, die übergeordnete Ebene meldet jedoch nur "Unbekannter Fehler". Systemverwalter sollten die Meldung "Unbekannter Fehler" aus der übergeordneten Ebene ignorieren.

Die folgenden Abschnitte enthalten DR-spezifische Erwägungen zu den unterschiedlichen Gerätetypen.

Erwägungen zur DR von CPU-Geräten im Cluster

Die Sun Cluster-Software wird einen DR-Vorgang zur Board-Entfernung aufgrund der vorhandenen CPU-Geräte nicht ablehnen.

Wenn ein DR-Vorgang zur Board-Hinzufügung erfolgt, werden die CPU-Geräte auf dem hinzugefügten Board automatisch in den Systembetrieb integriert.

Erwägungen zur DR von Speichern im Cluster

Für DR-Zwecke sind zwei Speichertypen zu berücksichtigen. Diese beiden Typen unterscheiden sich lediglich bei ihrer Nutzung. Die aktuelle Hardware ist bei beiden Typen die gleiche.

Der vom Betriebssystem genutzte Speicher wird als Kernel-Speichergehäuse bezeichnet. Die Sun Cluster-Software unterstützt die Board-Entfernung bei dem Board mit dem Kernel-Speichergehäuse nicht und lehnt jeden derartigen Vorgang ab. Wenn ein DR-Vorgang zur Board-Entfernung einen Speicher betrifft, der nicht das Kernel-Speichergehäuse ist, lehnt Sun Cluster den Vorgang nicht ab.

Wenn ein den Speicher betreffender DR-Vorgang zur Board-Hinzufügung durchgeführt wird, wird der Speicher auf dem hinzugefügten Board automatisch in den Systembetrieb integriert.

Erwägungen zur DR von Platten- und Bandlaufwerken im Cluster

Sun Cluster lehnt DR-Vorgänge zur Board-Entfernung auf aktiven Laufwerken des Primärknotens ab. DR-Vorgänge zur Board-Entfernung können auf nicht aktiven Laufwerken des Primärknotens und auf allen Laufwerken der Sekundärknoten durchgeführt werden. Der Cluster-Datenzugriff unterliegt durch den DR-Vorgang keinerlei Änderungen.

Hinweis – Sun Cluster lehnt DR-Vorgänge ab, die sich auf die Verfügbarkeit von Quorum-Geräten auswirken. Weitere Erwägungen zu Quorum-Geräten und dem Verfahren zur Ausführung von DR-Vorgängen bei Cluster-Geräten finden Sie unter „Erwägungen zur DR von Quorum-Geräten im Cluster“ auf Seite 89.

Detaillierte Anweisungen zur Ausführung dieser Aktionen finden Sie im *Sun Cluster 3.1 Handbuch Systemverwaltung*.

Erwägungen zur DR von Quorum-Geräten im Cluster

Wenn der DR-Vorgang zur Board-Entfernung ein Board betrifft, das eine Schnittstelle zu einem für das Quorum konfigurierten Gerät enthält, lehnt Sun Cluster den Vorgang ab und identifiziert das Quorum-Gerät, das von diesem Vorgang betroffen wäre. Sie müssen dieses Gerät als Quorum-Gerät deaktivieren, bevor Sie einen DR-Vorgang zur Board-Entfernung durchführen können.

Detaillierte Anweisungen zur Ausführung dieser Aktionen finden Sie im *Sun Cluster 3.1 Handbuch Systemverwaltung*.

Erwägungen zur DR von Cluster-Interconnect-Schnittstellen im Cluster

Wenn der DR-Vorgang zur Board-Entfernung ein Board betrifft, das zu einer aktiven Cluster-Interconnect-Schnittstelle gehört, lehnt Sun Cluster den Vorgang ab und identifiziert die Schnittstelle, die von dem Vorgang betroffen wäre. Sie müssen die aktive Schnittstelle mit einem Sun Cluster-Verwaltungstool deaktivieren, bevor der DR-Vorgang erfolgen kann (siehe auch den nachstehenden Abschnitt neben "Achtung").

Detaillierte Anweisungen zur Ausführung dieser Aktionen finden Sie im *Sun Cluster 3.1 Handbuch Systemverwaltung*.

Achtung – Sun Cluster erfordert, dass jeder Cluster-Knoten über mindestens einen funktionsfähigen Pfad zu jedem Cluster-Knoten verfügt. Deaktivieren Sie keine privaten Interconnect-Schnittstellen, die den letzten Pfad zu einem Cluster-Knoten unterstützen.

Erwägungen zur DR von öffentlichen Netzwerkschnittstellen im Cluster

Wenn der DR-Vorgang zur Board-Entfernung ein Board betrifft, das aktive öffentliche Netzwerkschnittstellen enthält, lehnt Sun Cluster den Vorgang ab und identifiziert die Schnittstelle, die von dem Vorgang betroffen wäre. Bevor ein Board mit einer aktiven öffentlichen Netzwerkschnittstelle entfernt werden kann, muss der gesamte Datenverkehr mithilfe des `if_mpadm`(1M)-Befehls von dieser Schnittstelle auf eine andere funktionsfähige Schnittstelle in der Multipathing-Gruppe umgeleitet werden.

Achtung – Wenn der verbleibende Netzwerkadapter während des DR-Entferngungsvorgangs für den deaktivierten Netzwerkadapter ausfällt, wird die Verfügbarkeit beeinträchtigt. Der verbleibende Adapter hat keine Möglichkeit, für die Dauer des DR-Vorgangs zu wechseln.

Detaillierte Anweisungen zur Durchführung eines DR-Vorgangs zur Entfernung einer öffentlichen Netzwerkschnittstelle finden Sie im *Sun Cluster 3.1 Handbuch Systemverwaltung*.

KAPITEL 4

Häufig gestellte Fragen

In diesem Kapitel finden Sie Antworten auf die am häufigsten gestellten Fragen zum SunPlex-System. Die Fragen sind nach Thema geordnet.

Häufig gestellte Fragen zur Hochverfügbarkeit

- **Was genau ist ein hoch verfügbares System?**

  Das SunPlex-System definiert Hochverfügbarkeit (HA) als die Fähigkeit eines Clusters, eine Anwendung auch dann aktiv und am Laufen zu halten, wenn ein Fehler aufgetreten ist, der normalerweise die Verfügbarkeit eines Serversystems unterbricht.

- **Mit welchem Prozess sorgt der Cluster für Hochverfügbarkeit?**

  Mit einem Prozess, der als Failover bezeichnet wird, sorgt das Cluster-Framework für eine hoch verfügbare Umgebung. Ein Failover ist eine Reihe von Schritten, mit denen der Cluster Datendienstressourcen von einem versagenden Knoten auf einen anderen funktionsfähigen Knoten im Cluster migriert.

- **Was ist der Unterschied zwischen einem Failover- und einem Scalable-Datendienst?**

  Es gibt zwei Typen von hoch verfügbaren Datendiensten, Failover und Scalable.

  Ein Failover-Datendienst führt eine Anwendung jeweils nur auf einem Primärknoten im Cluster aus. Andere Knoten können andere Anwendungen ausführen, aber jede Anwendung wird nur einmal auf einem einzigen Knoten ausgeführt. Wenn ein Primärknoten ausfällt, wird die auf dem ausgefallenen Knoten ausgeführte Anwendung mit einem Failover auf einen anderen Knoten verschoben und dort weiter ausgeführt.

Ein Scalable-Dienst verteilt eine Anwendung auf mehrere Knoten, um einen einzigen logischen Dienst zu erstellen. Scalable-Dienste setzen die Anzahl der Knoten und Prozessoren auf dem ganzen Cluster, auf dem sie durchgeführt werden, wirksam ein.

Für jede Anwendung hostet ein Knoten die reale Schnittstelle zum Cluster. Dieser Knoten wird als globaler Schnittstellenknoten (GIF-Knoten) bezeichnet. Im Cluster können mehrere GIF-Knoten vorhanden sein. Jeder GIF-Knoten hostet eine oder mehr logische Schnittstellen, die von Scalable-Diensten genutzt werden können. Diese logischen Schnittstellen werden als *globale Schnittstellen* bezeichnet. Ein GIF-Knoten hostet eine globale Schnittstelle für alle Anforderungen einer bestimmten Anwendung und sendet sie an mehrere Knoten, auf denen der Anwendungsserver läuft. Wenn der GIF-Knoten ausfällt, wird die globale Schnittstelle auf einen auch weiterhin laufenden Knoten umgeleitet.

Wenn einer der Knoten, auf denen die Anwendung ausgeführt wird, ausfällt, wird die Anwendung mit einem geringen Leistungsverlust auf anderen Knoten weiter ausgeführt, bis der ausgefallene Knoten wieder zum Cluster hinzukommt.

Häufige Fragen zu Dateisystemen

- **Kann mindestens ein Cluster-Knoten als hoch verfügbare(r) NFS-Server mit anderen Cluster-Knoten als Clients konfiguriert werden?**

  Nein, schaffen Sie keine Schleifeneinhängung.

- **Kann ein Cluster-Dateisystem für Anwendungen verwendet werden, die nicht von Ressourcengruppen-Manager gesteuert werden?**

  Ja. Allerdings müssen die Anwendungen ohne Steuerung durch den RGM nach dem Ausfall des Knotens, auf dem sie ausgeführt werden, manuell neu gestartet werden.

- **Müssen alle Cluster-Dateisysteme einen Einhängepunkt unter dem /global-Verzeichnis haben?**

  Nein. Aber Cluster-Dateisysteme mit demselben Einhängepunkt (wie zum Beispiel `/global`) ermöglichen eine bessere Organisation und Verwaltung dieser Dateisysteme.

- **Welches sind die Unterschiede zwischen der Verwendung eines Cluster-Dateisystems und dem Exportieren von NFS-Dateisystemen?**

  Es gibt mehrere Unterschiede:

  1. Das Cluster-Dateisystem unterstützt globale Geräte. NFS unterstützt keinen Remote-Zugriff auf Geräte.
  2. Das Cluster-Dateisystem hat einen globalen Namensraum. Nur ein einziger Einhängebefehl ist erforderlich. Bei NFS müssen Sie das Dateisystem auf jedem Knoten einhängen.

3. Das Cluster-Dateisystem speichert die Dateien in mehr Fällen als NFS im Cache. Dies geschieht zum Beispiel dann, wenn auf eine Datei von mehreren Knoten aus zum Lesen, Schreiben, für Dateisperren oder asynchronen E/A zugegriffen wird.
4. Das Cluster-Dateisystem unterstützt ein nahtloses Failover, wenn ein Server ausfällt. NFS unterstützt mehrere Server, aber ein Failover ist nur bei schreibgeschützten Dateisystemen möglich.
5. Das Cluster-Dateisystem ist auf die Nutzung zukünftiger schneller Cluster-Interconnects ausgelegt, die DMA-Funktionen und Funktionen ohne Zwischenspeicherung (Zero-Copy) im Fernzugriff zur Verfügung stellen.
6. Wenn Sie die Attribute für eine Datei (zum Beispiel mit `chmod(1M)`) in einem Cluster-Dateisystem ändern, wird diese Änderung sofort auf alle Knoten gespiegelt. Bei einem exportierten NFS-Dateisystem kann das wesentlich mehr Zeit in Anspruch nehmen.

- **Das Dateisystem /global/.devices/node@<Knoten-ID> wird auf den Cluster-Knoten angezeigt. Kann dieses Dateisystem zum Speichern von Daten verwendet werden, die hoch verfügbar und global sein sollen?**

  Diese Dateisysteme speichern den Namensraum für globale Geräte. Sie sind nicht zum allgemeinen Gebrauch vorgesehen. Sie sind zwar global, aber der Zugriff auf sie erfolgt niemals global -- jeder Knoten kann nur auf seinen eigenen Namensraum für globale Geräte zugreifen. Wenn ein Knoten ausgefallen ist, können andere Knoten nicht auf den Namensraum für den ausgefallenen Knoten zugreifen. Diese Dateisysteme sind nicht hoch verfügbar. Sie sollten nicht zum Speichern von Daten verwendet werden, die global zugänglich oder hoch verfügbar sein müssen.

Häufige Fragen zur Datenträgerverwaltung

- **Müssen alle Plattengeräte gespiegelt werden?**

  Damit ein Plattengerät als hoch verfügbar gilt, muss es gespiegelt werden oder RAID-5-Hardware verwenden. Alle Datendienste sollten entweder hoch verfügbare Plattengeräte oder Cluster-Dateisysteme verwenden, die auf hoch verfügbaren Plattengeräten eingehängt sind. Solche Konfigurationen tolerieren einzelne Plattenausfälle.

- **Kann ein Datenträger-Manager für die lokalen Platten (Boot-Platte) und ein anderer für die Multihostplatten verwendet werden?**

  Diese Konfiguration wird von der Solaris Volume Manager-Software zur Verwaltung der lokalen Platten und von VERITAS Volume Manager zur Verwaltung der Multihostplatten unterstützt. Keine andere Kombination wird

unterstützt.

Häufige Fragen zu Datendiensten

- **Welche SunPlex-Datendienste sind verfügbar?**

  Die Liste der unterstützten Datendienste finden Sie in *Sun Cluster 3.1 Versionshinweise*.

- **Welche Anwendungsversionen werden von den SunPlex-Datendiensten unterstützt?**

  Die Liste der unterstützten Anwendungsversionen ist in *Sun Cluster 3.1 Versionshinweise* enthalten.

- **Können eigene Datendienste geschrieben werden?**

  Ja. Weitere Informationen finden Sie im *Sun Cluster 3.1 Entwicklerhandbuch Datendienste* und in der Dokumentation zur Datendienst-Grundlagentechnologie, die mit der API für die Datendienst-Entwicklungsbibliothek mitgeliefert wird.

- **Müssen beim Erstellen von Netzwerkressourcen numerische IP-Adressen oder Hostnamen angegeben werden?**

  Die bevorzugte Methode zur Angabe von Netzwerkressourcen ist die Verwendung des UNIX-Hostnamens anstelle der numerischen IP-Adresse.

- **Was ist beim Erstellen von Netzwerkressourcen der Unterschied zwischen der Verwendung eines logischen Hostnamens (LogicalHostname-Ressource) und einer gemeinsam genutzten Adresse (SharedAddress-Ressource)?**

  Wenn in der Dokumentation die Verwendung einer `LogicalHostname`-Ressource in einer Ressourcengruppe im `Failover`-Modus beschrieben ist, kann wahlweise eine `SharedAddress`-Ressource oder eine `LogicalHostname`-Ressource verwendet werden. Die einzige Ausnahme bildet Sun Cluster HA für NFS. Die Verwendung einer `SharedAddress`-Ressource sorgt für zusätzlichen Überlauf, weil die Cluster-Netzwerksoftware für eine `SharedAddress`, aber nicht für einen `LogicalHostname` konfiguriert ist.

  Der Vorteil einer `SharedAddress` ergibt sich, wenn Sie sowohl Scalable- als auch Failover-Datendienste konfigurieren und möchten, dass Clients mithilfe desselben Hostnamens auf beide Dienste zugreifen können. In diesem Fall ist bzw. sind die `SharedAddress`-Ressource(n) zusammen mit der Failover-Anwendungsressource in einer Ressourcengruppe zusammengefasst, während die Scalable-Dienstressource in einer separaten Ressourcengruppe untergebracht und für die Verwendung der `SharedAddress` konfiguriert ist. Sowohl Scalable- als auch Failover-Dienste können denselben Satz von Hostnamen/Adressen verwenden, der in der `SharedAddress`-Ressource konfiguriert ist.

Häufige Fragen zum öffentlichem Netzwerk

- **Welche öffentlichen Netzwerkadapter werden vom SunPlex-System unterstützt?**

  Derzeit unterstützt das SunPlex-System öffentliche Netzwerkadapter für Ethernet (10/100BASE-T und 1000BASE-SX Gb). Zukünftig werden ggf. neue Schnittstellen unterstützt. Aktuelle Informationen erhalten Sie von Ihrem Sun-Vertreter.

- **Welche Rolle spielt die MAC-Adresse beim Failover?**

  Bei einem Failover werden neue ARP-Pakete (Address Resolution Protocol-Pakete) generiert und an alle gesendet. Diese ARP-Pakete enthalten die neue MAC-Adresse (des neuen tatsächlichen Adapters, zu dem der Knoten gewechselt ist) und die alte IP-Adresse. Wenn ein anderer Rechner im Netzwerk eines dieser Pakete erhält, löscht er die alte MAC-IP-Zuordnung aus seinem ARP-Cache und verwendet die neue.

- **Unterstützt das SunPlex-System die Einstellung local-mac-address?=true im OpenBoot™ PROM (OBP) für einen Hostadapter?**

  Ja Tatsächlich erfordert IP Network Multipathing, dass `local-mac-address?` auf `true` gesetzt sein muss.

- **Welche Verzögerung ist zu erwarten, wenn IP Network Multipathing ein Switchover zwischen Adaptern ausführt?**

  Die Verzögerung kann mehrere Minuten betragen. Der Grund dafür ist, dass ein IP Network Multipathing-Switchover das Senden eines freien ARP beinhaltet. Es gibt jedoch keine Garantie dafür, dass der Router zwischen Client und Cluster dieses ARP auch verwendet. Bis zum Ablauf der Gültigkeit des ARP-Cache-Eintrags für diese IP-Adresse auf dem Router wird möglicherweise die veraltete MAC-Adresse verwendet.

- **Wie schnell werden Fehler an einem Netzwerkadapter erkannt?**

  Die Standard-Fehlererkennungszeit beträgt 10 Sekunden. Der Algorithmus versucht, die Fehlererkennungszeit einzuhalten, doch die tatsächlich benötigte Zeit hängt von der Netzwerklast ab.

Häufige Fragen zu Cluster-Mitgliedern

- **Muss das Root-Passwort für alle Cluster-Mitglieder gleich sein?**

  Sie sind nicht gezwungen, dasselbe Root-Passwort für jedes Cluster-Mitglied zu verwenden. Sie können jedoch die Cluster-Verwaltung vereinfachen, wenn Sie auf allen Knoten dasselbe Root-Passwort verwenden.

- **Ist es wichtig, in welcher Reihenfolge die Knoten gebootet werden?**

  In den meisten Fällen nicht. Die Boot-Reihenfolge ist jedoch wichtig, um Amnesie zu verhindern (Details zu Amnesie finden Sie unter „Quorum und Quorum-Geräte“ auf Seite 54). Wenn Knoten Zwei zum Beispiel der Eigentümer des Quorum-Geräts war, Knoten Eins nicht aktiv ist und Sie dann Knoten Zwei herunterfahren, müssen Sie Knoten Zwei vor Knoten Eins wieder hochfahren. Damit wird verhindert, dass Sie versehentlich einen Knoten mit veralteten Cluster-Konfigurationsinformationen starten.

- **Müssen lokale Platten in einem Cluster-Knoten gespiegelt werden?**

  Ja Diese Spiegelung ist zwar keine Voraussetzung, aber das Spiegeln der Platten der Cluster-Knoten beugt einem Ausfall einer nicht gespiegelten Platte vor, der einen Knotenausfall nach sich zieht. Der Nachteil beim Spiegeln lokaler Platten eines Cluster-Knotens ist ein erhöhter Systemverwaltungsaufwand.

- **Welche Sicherungslösungen gibt es für Cluster-Mitglieder?**

  Sie können mehrere Sicherungsmethoden für einen Cluster verwenden. Eine Methode besteht darin, einen Knoten als Sicherungsknoten mit angeschlossenem Bandlaufwerk/Bibliothek zu verwenden. Dann sichern Sie die Daten mit dem Cluster-Dateisystem. Verbinden Sie diesen Knoten nicht mit den gemeinsam genutzten Platten.

  Weitere Informationen zu Sicherungs- und Wiederherstellungsverfahren finden Sie im *Sun Cluster 3.1 Handbuch Systemverwaltung*.

- **Wann ist ein Knoten stabil genug, um als Sekundärknoten eingesetzt zu werden?**

  Nach dem erneuten Booten ist ein Knoten stabil genug, um als Sekundärknoten eingesetzt zu werden, wenn der Knoten die Anmelde-Eingabeaufforderung anzeigt.

Häufige Fragen zum Cluster-Speicher

- **Was macht den Multihost-Speicher hoch verfügbar?**

  Multihost-Speicher ist hoch verfügbar, weil er den Verlust einer Einzelplatte dank der Spiegelung (oder dank hardwarebasierter RAID-5-Controller) übersteht. Ein Multihost-Speichergerät hat mehr als eine Hostverbindung. Deswegen widersteht es dem Verlust eines einzelnen Knotens, an den es angeschlossen ist. Zudem sorgen redundante Pfade von jedem Knoten zum angeschlossenen Speicher dafür, dass ein Versagen eines Hostbusadapters, Kabels oder Platten-Controllers toleriert wird.

Häufige Fragen zum Cluster-Interconnect

- **Welche Cluster-Interconnect werden vom SunPlex-System unterstützt?**

  Derzeit unterstützt das SunPlex-System die Cluster-Interconnect für Ethernet (100BASE-T Fast Ethernet und 1000BASE-SX Gb) und für die SCI-Netzwerkschnittstelle.

- **Was ist der Unterschied zwischen einem "Kabel" und einem "Transportpfad"**

  Cluster-Transportkabel werden unter Verwendung von Transportadaptern und Schaltern konfiguriert. Kabel verbinden Adapter und Schalter auf einer Komponente-zu-Komponenten-Basis. Der Cluster-Topologie-Manager setzt die verfügbaren Kabel ein, um durchgehende (von Ende zu Ende) Transportpfade zwischen den Knoten aufzubauen. Ein Kabel ist einem Transportpfad nicht direkt zugeordnet.

  Kabel werden von einem Verwalter statisch "aktiviert" und "deaktiviert". Bei Kabeln gibt es einen "Zustand", (aktiv oder inaktiv), aber keinen "Status". Ein inaktives Kabel ist wie ein nicht konfiguriertes Kabel. Inaktive Kabel können nicht als Transportpfade eingesetzt werden. Sie werden nicht getestet, und deswegen ist ihr Status nicht bekannt. Der Zustand eines Kabels kann mit `scconf -p` angezeigt werden.

  Transportpfade werden dynamisch durch den Cluster-Topologie-Manager festgelegt. Der "Status" eines Transportpfades wird vom Topologie-Manager bestimmt. Ein Pfad kann den Status "online" oder "offline" haben. Der Status eines Transportpfades kann mit `scstat(1M)` angezeigt werden.

  Betrachten Sie das folgende Beispiel eines Zwei-Knoten-Clusters mit vier Kabeln.

  ```
  Knoten1:Adapter0     an Schalter1, Port0
  Knoten1:Adapter1     an Schalter2, Port0
  Knoten2:Adapter0     an Schalter1, Port1
  Knoten2:Adapter1     an Schalter2, Port1
  ```

  Es gibt zwei mögliche Transportpfade, die aus diesen vier Kabeln gebildet werden können.

  ```
  Knoten1:Adapter0    an Knoten2:Adapter0
  Knoten2:Adapter1    an
  Knoten2:Adapter1
  ```

Häufige Fragen zu Client-Systemen

- **Müssen bei der Verwendung eines Clusters besondere Client-Anforderungen oder Einschränkungen berücksichtigt werden?**

  Client-Systeme stellen mit dem Cluster die Verbindung auf dieselbe Weise wie mit einem anderen Server her. Bei manchen Instanzen kann es je nach Datendienstanwendung erforderlich sein, clientseitige Software zu installieren oder sonstige Änderungen an der Konfiguration vorzunehmen, damit der Client eine Verbindung mit der Datendienstanwendung herstellen kann. Weitere Informationen zu clientseitigen Konfigurationsanforderungen finden Sie im *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Häufige Fragen zur Verwaltungskonsole

- **Erfordert das SunPlex-System eine Verwaltungskonsole?**

  Ja.

- **Muss die Verwaltungskonsole dem Cluster zugewiesen werden oder kann sie auch für andere Aufgaben eingesetzt werden?**

  Das SunPlex-System erfordert keine dedizierte Verwaltungskonsole, aber diese hat folgende Vorteile:

  - Sie ermöglicht eine zentralisierte Cluster-Verwaltung durch das Gruppieren von Konsolen- und Verwaltungstools auf demselben Rechner.
  - Sie sorgt für eine potenziell schnellere Problemlösung durch Ihren Hardware-Dienstleister.

- **Muss sich die Verwaltungskonsole in der "Nähe" des Clusters befinden, zum Beispiel in demselben Raum?**

  Fragen Sie bei Ihrem Hardware-Dienstleister nach. Der Dienstleister kann ggf. die räumliche Nähe der Konsole zum eigentlichen Cluster vorschreiben. Es gibt keine technischen Gründe, die eine Unterbringung der Konsole in demselben Raum erforderlich machen.

- **Können Verwaltungskonsolen mehr als einen Cluster verwalten, wenn ggf. vorhandene Abstandsvorgaben zuvor eingehalten werden?**

  Ja. Sie können von einer einzigen Verwaltungskonsole aus mehrere Cluster steuern. Sie können auch einen einzigen Terminal-Konzentrator zwischen den Clustern teilen.

Häufige Fragen zu Terminal-Konzentrator und System Service Processor

- **Erfordert das SunPlex-System einen Terminal-Konzentrator?**

  Softwareversionen ab Sun Cluster 3.0 benötigen keinen Terminal-Konzentrator. Anders als das Produkt Sun Cluster 2.2, das einen Terminal-Konzentrator als Fehlerschutz benötigte, hängen neuere Produkte nicht vom Terminal-Konzentrator ab.

- **Die meisten SunPlex-Server verwenden einen Terminal-Konzentrator, aber Sun Enterprise E10000-Server nicht. Warum?**

  Der Terminal-Konzentrator ist für die meisten Server tatsächlich ein Seriell-zu-Ethernet-Konverter. Sein Konsolen-Port ist ein serieller Port. Der Sun Enterprise E10000-Server hat keine serielle Konsole. Der System Service Processor (SSP) ist die Konsole, entweder über einen Ethernet- oder einen `jtag`-Port. Bei Sun Enterprise E10000-Server verwenden Sie immer den SSP als Konsole.

- **Welches sind die Vorteile eines Terminal-Konzentrators?**

  Mit einem Terminal-Konzentrator können Sie auf Konsolenebene von einer Remote-Workstation an einer beliebigen Stelle im Netzwerk selbst dann auf jeden Knoten zugreifen, wenn sich der Knoten am OpenBoot PROM (OBP) befindet.

- **Welches sind die für die Qualifizierung eines nicht von Sun unterstützten Terminal-Konzentrators erforderlichen Kenntnisse?**

  Der Hauptunterschied zwischen dem von Sun unterstützen Terminal-Konzentrator und anderen Konsolengeräten besteht darin, dass der Terminal-Konzentrator von Sun mit spezieller Firmware ausgestattet ist, die ihn daran hindert, einen Abbruch an die bootende Konsole zu senden. Beachten Sie, dass der Knoten durch ein Konsolengerät heruntergefahren wird, wenn das Gerät einen Abbruch oder ein von der Konsole als Abbruch interpretiertes Signal senden kann.

- **Kann ein gesperrter Port am von Sun unterstützten Terminal-Konzentrator freigeschaltet werden, ohne ihn neu zu booten?**

  Ja. Beachten Sie, dass die Port-Nummer neu eingestellt werden muss, und führen Sie Folgendes aus:

```
telnet tc
Geben Sie den Annex-Port-Namen oder -Nummer ein: cli
annex: su -
annex# admin
admin : reset Port-Nummer
admin : quit
annex# hangup
#
```

Weitere Informationen zur Konfiguration und Verwaltung des von Sun unterstützten Terminal-Konzentrators finden Sie im *Sun Cluster 3.1 Handbuch Systemverwaltung*.

- **Was geschieht, wenn der Terminal-Konzentrator selbst ausfällt? Ist ein zweiter als Standby-Gerät erforderlich?**

  Nein. Bei einem Versagen des Terminal-Konzentrators geht keine Cluster-Verfügbarkeit verloren. Sie können erst dann wieder eine Verbindung mit den Knotenkonsolen herstellen, wenn der Konzentrator wieder arbeitet.

- **Wenn sieht es bei einem Terminal-Konzentrator in puncto Sicherheit aus?**

  Im Allgemeinen ist der Terminal-Konzentrator mit einem kleinen Netzwerk verbunden, das von Systemverwaltern verwendet wird, und nicht mit einem Netzwerk, das für sonstigen Client-Zugriff eingesetzt wird. Sie können die Sicherheit durch einen beschränkten Zugriff auf dieses Netzwerk steuern.

- **Wie wird die dynamische Rekonfiguration für ein Band- oder Plattenlaufwerk eingesetzt?**

  - Stellen Sie fest, ob das Platten- oder Bandlaufwerk zu einer aktiven Gerätegruppe gehört. Wenn das Laufwerk zu keiner aktiven Gerätegruppe gehört, können Sie den DR-Entfernungsvorgang ausführen.
  - Wenn der DR-Vorgang zur Board-Entfernung ein aktives Platten- oder Bandlaufwerk betrifft, lehnt das System den Vorgang ab und identifiziert die betroffenen Laufwerke. Wenn das Laufwerk zu einer aktiven Gerätegruppe gehört, gehen Sie zu „Erwägungen zur DR von Platten- und Bandlaufwerken im Cluster“ auf Seite 89.
  - Stellen Sie fest, ob das Laufwerk eine Komponente des Primär- oder Sekundärknotens ist. Wenn das Laufwerk eine Komponente des Sekundärknotens ist , können Sie den DR-Entfernungsvorgang ausführen.
  - Wenn das Laufwerk eine Komponente des Primärknotens ist, müssen Sie vom Primär- auf den Sekundärknoten umschalten, bevor Sie den DR-Entfernungsvorgang für das Gerät durchführen können.

Achtung – Wenn der aktuelle Primärknoten ausfällt, während Sie den DR-Vorgang auf einem Sekundärknoten ausführen, wirkt sich das auf die Cluster-Verfügbarkeit aus. Der Primärknoten kann kein Failover auf einen anderen Knoten ausführen, bis ein neuer Sekundärknoten bereitgestellt wird.

Glossar

Dieses Begriffsglossar wird in der SunPlex 3.1-Dokumentation verwendet.

A

Amnesie Ein Zustand, bei dem der Cluster nach dem Herunterfahren mit veralteten Cluster-Konfigurationsdaten neu startet (CCR). Ist in einem Zwei-Cluster-Knoten zum Beispiel nur Knoten 1 funktionsfähig, sind nach einer Änderung der Cluster-Konfiguration auf Knoten 1 die CCR für Knoten 2 veraltet. Wenn der Cluster heruntergefahren und auf Knoten 2 neu gestartet wird, kommt es aufgrund der veralteten CCR von Knoten 2 zu einem Zustand der Amnesie.

Automatisches Failback Der Prozess, mit dem eine Ressourcen- oder Gerätegruppe an den Primärknoten zurückgegeben wird, nachdem der Primärknoten versagt hatte und als Cluster-Mitglied neu gestartet wurde.

C

Checkpoint Die Benachrichtigung, die von einem Primärknoten an einen Sekundärknoten gesendet wird, um die Software zwischen beiden synchron zu halten. Siehe auch “Primärknoten” und “Sekundärknoten”.

Cluster-Konfigurations-Repository (Cluster Configuration Repository, CCR)	Ein hoch verfügbarer, replizierter Datenspeicher, der von der Sun Cluster-Software zur dauerhaften Speicherung der Cluster-Konfigurationsinformationen verwendet wird.
Cluster-Dateisystem	Ein Cluster-Dienst, der Cluster-weiten, hoch verfügbaren Zugriff auf vorhandene lokale Dateisysteme bereitstellt.
Cluster-Interconnect	Die Hardware-Netzwerkinfrastruktur einschließlich Kabel, Cluster-Transportverbindungspunkte und Cluster-Transportadapter. Die Sun Cluster- und Datendienstsoftware nutzt diese Infrastruktur für die Cluster-interne Kommunikation.
Cluster-Knoten	Ein Knoten, der als Cluster-Mitglied konfiguriert ist. Ein Cluster-Knoten kann ein aktuelles Mitglied sein. Siehe auch “Cluster-Mitglied”.
Cluster-Mitglied	Ein aktives Mitglied der aktuellen Cluster-Zusammensetzung. Dieses Mitglied kann mit anderen Cluster-Mitgliedern Ressourcen teilen und Dienste sowohl für andere Cluster-Mitglieder als auch für Clients des Clusters zur Verfügung zu stellen. Siehe auch “Cluster-Knoten”.
Cluster-Mitglieder-Monitor (Cluster Membership Monitor, CMM)	Die Software, die eine ständige Überwachung der Cluster-Mitgliedschaft ausführt. Diese Mitgliederinformationen werden von allen übrigen Cluster-Softwareanwendungen verwendet, um zu entscheiden, wo hoch verfügbare Dienste ausgeführt werden sollen. CMM stellt sicher, dass keine Nicht-Cluster-Mitglieder Daten beschädigen bzw. beschädigte oder nicht konsistente Daten an Clients übertragen können.
Cluster-Transportadapter	Der Netzwerkadapter, der auf einem Knoten sitzt und den Knoten mit dem Cluster-Interconnect verbindet. Siehe auch “Cluster-Interconnect”.
Cluster-Transportkabel	Die Netzwerkverbindung, die die Verbindung mit den Endpunkten herstellt. Eine Verbindung zwischen Cluster-Transportadaptern und Cluster-Transportverbindungspunkten oder zwischen zwei Cluster-Transportadaptern. Siehe auch “Cluster-Interconnect”.
Cluster-Transportverbindungspunkt	Ein Hardwareschalter, der als Teil des Cluster-Interconnect verwendet wird. Siehe auch “Cluster-Interconnect”.
Cluster	Mindestens zwei miteinander verbundene Knoten oder Domänen mit einem gemeinsam genutzten Cluster-Dateisystem, die zusammen konfiguriert sind, um Failover-, Parallel- oder Scalable-Ressourcen auszuführen.

D

Datenbankreplikat Metagerätezustand (Replikat) Eine auf der Platte gespeicherte Datenbank, in der die Konfiguration und der Zustand aller Metageräte sowie die Fehlerbedingungen aufgezeichnet sind. Diese Informationen sind für einen ordnungsgemäßen Betrieb der Solaris Volume Manager-Plattensätze wichtig und werden repliziert.

Datendienst Eine Anwendung, die als hoch verfügbare Ressource unter der Steuerung von Ressourcengruppen-Manager (RGM) eingerichtet wurde.

Datenträger-Manager Ein Softwareprodukt, das mithilfe von Disk Striping, Verkettung, Spiegelung und dynamischem Wachstum von Metageräten oder Datenträgern für die Zuverlässigkeit von Daten sorgt.

DID-Name Wird zur Identifizierung von globalen Geräten auf einem SunPlex-System verwendet. Dabei handelt es sich um eine Cluster-Kennung mit einer 1:1- oder 1:M-Beziehung mit logischen Solaris-Namen. Der Name hat die Form d*X*s*Y*, wobei *X* eine ganze Zahl und *Y* der Bereichsname ist. Siehe auch "Logischer Solaris-Name".

DID-Treiber Ein Treiber, der von der Sun Cluster-Software implementiert wird und zur Bereitstellung eines konsistenten Geräte-Namensraums im ganzen Cluster dient. Siehe auch "DID-Name".

Distributed Lock Manager (DLM) Die Sperrsoftware, die in einer gemeinsam genutzten OPS-Plattenumgebung (Oracle Parallel Server) verwendet wird. DLM ermöglicht das Ausführen von Oracle-Prozessen auf verschiedenen Knoten zum Synchronisieren des Datenbankzugriffs. DLM ist auf hohe Verfügbarkeit ausgelegt. Wenn ein Prozess oder ein Knoten abstürzt, müssen die restlichen Knoten nicht heruntergefahren und neu gestartet werden. Es wird eine rasche Rekonfiguration von DLM ausgeführt, um einen solchen Ausfall zu beheben.

E

Einzelinstanz-Ressource Eine Ressource, für die maximal eine Ressource auf dem ganzen Cluster aktiv sein kann.

Endpunkt Ein realer Port auf einem Cluster-Transportadapter bzw. einem Cluster-Transportverbindungspunkt.

Ereignis Eine Änderung bei Zustand, Beherrschung, Schwere oder Beschreibung eines verwalteten Objekts.

F

Failback Siehe “Automatisches Failback”.

Failfast Das ordnungsgemäße Herunterfahren eines fehlerhaften Knotens und Entfernen vom Cluster, bevor dessen möglicherweise fehlerhafter Betrieb Schaden anrichten kann.

Failover-Ressource Eine Ressource, deren Ressourcen alle jeweils nur von einem einzigen Knoten richtig unterstützt werden können. Siehe auch “Einzelinstanz-Ressource” und “Scalable-Ressource”.

Failover Die automatische Neuzuweisung einer Ressourcengruppe oder Gerätegruppe von einem aktuellen Primärknoten zu einem neuen Primärknoten, nachdem ein Fehler aufgetreten ist.

Fehlerüberwachung Ein Fehler-Dämon und die Programme, die zur Überprüfung unterschiedlicher Teile von Datendiensten verwendet werden und Maßnahmen ergreifen. Siehe auch “Ressourcen-Monitor”.

G

Gemeinsamer Speicherort Die Eigenschaft, sich auf dem gleichen Knoten zu befinden. Dieses Konzept wird bei der Cluster-Konfiguration zur Leistungssteigerung eingesetzt.

Generische Ressource Ein Anwendungs-Dämon und dessen untergeordnete Prozesse, die als Teil eines generischen Ressourcentyps unter die Steuerung des Ressourcengruppen-Managers gestellt werden.

Generischer Ressourcentyp Eine Vorlage für einen Datendienst. Ein generischer Ressourcentyp kann verwendet werden, um eine einfache Anwendung zu einem Failover-Datendienst zu machen (Stoppen auf einem Knoten, Starten auf einem anderen Knoten). Dieser Typ muss nicht mit der SunPlex-API programmiert werden.

Geräte-ID

Ein Mechanismus zum Identifizieren von Geräten, die über Solaris bereitgestellt werden. Geräte-IDs werden in der Online-Dokumentation unter `devid_get(3DEVID)` beschrieben.

Der DID-Treiber von Sun Cluster legt mit Geräte-IDs den Bezug zwischen den logischen Solaris-Namen auf verschiedenen Cluster-Knoten fest. Der DID-Treiber sucht bei jedem Gerät nach der Geräte-ID. Wenn diese Geräte-ID mit einem anderen Gerät auf dem Cluster übereinstimmt, erhalten beide Geräte denselben DID-Namen. Wenn die Geräte-ID auf dem Cluster noch nicht vorhanden ist, wird ein neuer DID-Name zugewiesen. Siehe auch “Logischer Solaris-Name” und “DID-Treiber”.

Gerätegruppe

Eine benutzerdefinierte Gruppe von Geräteressourcen, wie zum Beispiel Platten, die von verschiedenen Knoten in einer hoch verfügbaren Cluster-Konfiguration unterstützt werden können. Diese Gruppe kann Geräteressourcen wie Platten, Solaris Volume Manager-Plattensätze und VERITAS Volume Manager-Plattengruppen enthalten.

GIF-Knoten

Siehe ”Globaler Schnittstellenknoten”.

Globale Ressource

Eine hoch verfügbare Ressource, die auf Kernel-Ebene der Sun Cluster-Software bereitgestellt wird. Globale Ressourcen umfassen Platten (HA-Gerätegruppen), das Cluster-Dateisystem und das globale Netzwerk.

Globale Schnittstelle

Eine globale Netzwerkschnittstelle, an der gemeinsam genutzte Adressen real gehostet werden. Siehe auch “Gemeinsam genutzte Adresse”.

Globaler Schnittstellenknoten (GIF-Knoten)

Ein Knoten, der eine globale Schnittstelle hostet.

Globales Gerät

Ein Gerät, das für alle Cluster-Mitglieder zugänglich ist, wie Platte, CD-ROM und Band.

H

HA-Datendienst

Siehe “Datendienst”.

Heartbeat

Eine über alle verfügbaren Transportpfade im Cluster-Interconnect regelmäßig ausgegebene Meldung. Das Ausbleiben eines Heartbeat nach einem festgelegten Zeitraum und einer festgelegten Anzahl von Versuchen löst ein internes Failover der Kommunikationsübertragung

auf einen anderen Pfad aus. Bei einem Versagen aller Pfade zu einem Cluster-Mitglied führt der Cluster-Mitglieder-Monitor (CMM) eine Neubewertung des Cluster-Quorums durch.

I

Instanz — Siehe “Ressourcenaufruf”.

Internet Protocol (IP) Network Multipathing — Diese Software verhindert durch Fehlerüberwachung und Failover einen Verlust der Knotenverfügbarkeit aufgrund des Ausfall eines einzelnen Netzwerkadapters oder Kabels. IP Network Multipathing-Failover setzt eine Reihe von Netzwerkadaptern mit der Bezeichnung IPMP-Gruppen (IP Network Multipathing-Gruppen) ein, um redundante Verbindungen zwischen einem Cluster-Knoten und dem öffentlichen Netzwerk bereitzustellen. Die Fehlerüberwachungs- und Failover-Funktionen arbeiten zusammen, um die Verfügbarkeit von Ressourcen sicherzustellen. Siehe auch “IPMP-Gruppe”.

IPMP-Gruppe (IP Network Multipathing-Gruppe) — Ein Satz aus einem oder mehreren Netzwerkadaptern auf demselben Knoten und in demselben Teilnetz, die so konfiguriert sind, dass sie im Fall eines Adapterversagens eine Sicherungskopie voneinander erstellen.

IPv4 — Internet Protocol, Version 4. Manchmal als IP bezeichnet. Diese Version unterstützt einen 32–Bit-Adressraum.

K

Knoten — Ein realer Rechner oder eine Domäne (im Sun Enterprise E10000-Server), das bzw. die zum SunPlex-System gehören kann. Auch als “Host” bezeichnet.

L

Lastausgleich Betrifft nur Scalable-Dienste. Der Prozess der Verteilung der Anwendungslast auf die im Cluster vorhandenen Knoten, um Client-Anforderungen zeitgerecht abzuwickeln. Weitere Details finden Sie unter „Scalable-Datendienste“ auf Seite 63.

Lastausgleichsverfahren Betrifft nur Scalable-Dienste. Die bevorzugte Form der Verteilung von Lastanforderungen aus der Anwendung auf die Knoten. Weitere Details finden Sie unter „Scalable-Datendienste“ auf Seite 63.

Laufende Aufrüstung In einer Sun Cluster-Konfiguration eine Aufrüstung, die sequenziell auf einem Cluster-Knoten nach dem anderen durchgeführt wird. Während einer laufenden Aufrüstung bleibt der Cluster in Betrieb und die Dienste werden auf den anderen Knoten weiter ausgeführt.

Logische Netzwerkschnittstelle In der Internet-Architektur kann ein Host eine oder mehrere IP-Adressen haben. Die Sun Cluster-Software konfiguriert zusätzliche logische Netzwerkschnittstellen, um eine Zuordnung zwischen mehreren logischen Netzwerkschnittstellen und einer einzelnen realen Netzwerkschnittstelle festzulegen. Jede logische Netzwerkschnittstelle hat eine einzige IP-Adresse. Durch diese Zuordnung kann eine einzige reale Netzwerkschnittstelle mehrere IP-Adressen abdecken. Durch diese Zuordnung kann zudem die IP-Adresse ohne zusätzliche Hardware-Schnittstellen bei Takeover- oder Switchover-Vorgängen von einem Cluster-Mitglied zu einem anderen verschoben werden.

Logischer Host Ein Konzept aus Sun Cluster 2.0 (mindestens), das eine Anwendung, die Plattensätze oder die Plattengruppen, auf denen sich die Anwendungsdaten befinden, und die Netzwerkadressen für den Zugriff auf den Cluster umfasst. Dieses Konzept ist im SunPlex-System nicht mehr vorhanden. Eine Beschreibung der aktuellen Implementierung dieses Konzepts im SunPlex-System finden Sie unter „Plattengerätegruppen“ auf Seite 42 und „Ressourcen, Ressourcengruppen und Ressourcentypen“ auf Seite 72.

Logischer Solaris-Name Die Namen, die in der Regel zur Verwaltung der Solaris-Geräte verwendet werden. Für Platten sehen diese in der Regel folgendermaßen aus: `/dev/rdsk/c0t2d0s2`. Zu jedem dieser logischen Solaris-Gerätenamen gibt es einen realen Solaris-Gerätenamen. Siehe auch “DID-Name” und “Realer Solaris-Name”.

Lokale Platte Eine Platte, die real ausschließlich zu einem gegebenen Cluster-Knoten gehört.

M

Master Siehe “Primärknoten”.

Multihomed Host Ein Host, der mit mehreren öffentlichen Netzwerken verbunden ist.

Multihostplatte Eine Platte, die real mit mehreren Knoten verbunden ist.

N

Namensraum globales Gerät Ein Namensraum mit den logischen, Cluster-weiten Namen für globale Geräte. Lokale Geräte sind in der Solaris-Umgebung in den Verzeichnissen `/dev/dsk`, `/dev/rdsk` und `/dev/rmt` definiert. Der Namensraum globales Gerät definiert globale Geräte in den Verzeichnissen `/dev/global/dsk`, `/dev/global/rdsk` und `/dev/global/rmt`.

Netzwerkadressressource Siehe “Netzwerkressource”.

Netzwerkressource Eine Ressource mit einem oder mehreren logischen Hostnamen oder gemeinsam genutzten Adressen. Siehe auch “Ressource logischer Hostname” und “Ressource gemeinsam genutzte Adresse”.

Nicht-Cluster-Modus Der Zustand nach dem Booten eines Cluster-Mitglieds mit der `-x`-Bootoption. In diesem Zustand ist der Knoten kein Cluster-Mitglied mehr, aber immer noch ein Cluster-Knoten. Siehe auch “Cluster-Mitglied” und “Cluster-Knoten”.

P

Parallele Dienstinstanz Eine Instanz für einen parallelen Ressourcentyp, die auf einem einzelnen Knoten läuft.

Paralleler Ressourcentyp Ein Ressourcentyp, zum Beispiel eine parallele Datenbank, die für den Betrieb in einer Cluster-Umgebung eingerichtet wurde, so dass sie von mehreren (zwei oder mehr) Knoten gleichzeitig unterstützt werden kann.

Plattengruppe Siehe “Gerätegruppe”.

Plattengerätegruppe Siehe “Gerätegruppe”.

Plattensatz Siehe “Gerätegruppe”.

Potenzieller Master Siehe “Potenzieller Primärknoten”.

Potenzieller Primärknoten Ein Cluster-Mitglied, das bei Versagen des Primärknotens in der Lage ist, einen Failover-Ressourcentyp zu unterstützen. Siehe auch “Standard-Master”.

Primärhostname Der Name eines Knotens im primären öffentlichen Netzwerk. Das ist immer der in `/etc/nodename` angegebene Knotennamen. Siehe auch “Sekundärhostname”.

Primärknoten Ein Knoten, auf dem eine Ressourcengruppe oder eine Gerätegruppe derzeit online ist. Ein Primärknoten ist mit anderen Worten ein Knoten, der derzeit die der Ressource zugeordneten Dienste hostet oder ausführt. Siehe auch “Sekundärknoten”.

Privater Hostname Der Alias für den Hostnamen, der für die Kommunikation mit einem Knoten über den Cluster-Interconnect verwendet wird.

Q

Quorum-Gerät Eine Platte, die von zwei oder mehr Knoten gemeinsam genutzt wird und Stimmen abgibt. Die Stimmen dienen der Feststellung des Quorums für den Betrieb des Clusters. Der Cluster kann nur arbeiten, wenn ein Quorum von Stimmen verfügbar ist. Das Quorum-Gerät wird verwendet, wenn ein Cluster in separate Knotensätze partitioniert wird, um festzulegen, welcher Knotensatz den neuen Cluster bildet.

R

Realer Solaris-Name Der Name, den das Gerät vom dazugehörigen Gerätetreiber in Solaris erhält. Er wird auf einem Solaris-Rechner als Pfad im Verzeichnisbaum unter `/devices` angezeigt. Der reale Solaris-Name für ein typisches SCSI-Laufwerk sieht zum Beispiel etwa folgendermaßen aus:`/devices/sbus@1f,0/SUNW,fas@e,8800000/sd@6,0:c,raw`

Siehe auch “Logischer Solaris-Name”.

Ressource gemeinsam genutzte Adresse Eine Netzwerkadresse, die von allen auf den Knoten innerhalb des Clusters laufenden Scalable-Diensten eingesetzt werden kann, um die Dienste auf diesen Knoten zu skalieren. Ein Cluster kann mehrere gemeinsam genutzte Adressen haben, und ein Dienst kann an mehrere gemeinsam genutzte Adressen angebunden sein.

Ressource logischer Hostname Eine Ressource mit einer Sammlung logischer Hostnamen, welche die Netzwerkadressen darstellen. Diese Ressourcen können nur jeweils von einem Knoten unterstützt werden. Siehe auch "Logischer Host".

Ressource Eine Instanz eines Ressourcentyps. Es können viele Ressourcen desselben Typs vorhanden sein, jede Ressource mit eigenem Namen und eigenen Eigenschaftswerten, so dass viele Instanzen der zugrunde liegenden Anwendung auf dem Cluster ausgeführt werden können.

Ressourcen-Monitor Ein optionaler Bestandteil einer Ressourcentyp-Implementierung, der regelmäßig Fehlerprüfungen bei Ressourcen durchführt, um festzustellen, ob diese ordnungsgemäß laufen und wie sich ihre Leistung darstellt.

Ressourcenaufruf Eine Instanz eines Ressourcentyps, die auf einem Knoten ausgeführt wird. Es handelt sich um ein abstraktes Konzept zur Darstellung einer Ressource, die auf dem Knoten gestartet wurde.

Ressourcengruppe Eine Sammlung von Ressourcen, die vom RGM als Einheit verwaltet werden. Jede Ressource, die vom RGM verwaltet werden soll, muss in einer Ressourcengruppe konfiguriert sein. In der Regel werden verwandte und voneinander abhängige Ressourcen in einer Gruppe zusammengefasst.

Ressourcengruppen-Manager (RGM) Ein Software-Leistungsmerkmal, mit dem Cluster-Ressourcen durch das automatische Starten und Anhalten der Ressourcen auf ausgewählten Cluster-Knoten hoch verfügbar und skalierbar gemacht werden. Der RGM agiert anhand vorkonfigurierter Verfahren im Fall von Hardware- oder Softwarefehlern oder beim erneuten Booten.

Ressourcengruppenzustand Der Zustand der Ressourcengruppe auf jedem Knoten.

Ressourcenstatus Der Status der Ressource, der vom Fehler-Monitor gemeldet wird.

Ressourcenzustand Der Zustand einer Ressourcengruppen-Manager-Ressource auf einem bestimmten Knoten.

Ressourcentyp Der einmalige Name eines Cluster-Objekts Datendienst, logischer Hostname oder gemeinsam genutzte Adresse. Ressourcentypen für Datendienste können vom Typ Failover oder Scalable sein. Siehe auch "Datendienst", "Failover-Ressource" und "Scalable-Ressource".

Ressourcentypeigenschaften Ein Paar von Schlüsselwerten, die vom Ressourcengruppen-Manager (RGM) als Teil des Ressourcentyps gespeichert und zur Beschreibung und Verwaltung der Ressourcen des angegebenen Typs verwendet werden.

Ressourcenverwaltungs-API (RMAPI) Die Anwendungsprogrammierschnittstelle (API) in einem SunPlex-System, die eine Anwendung in einer Cluster-Umgebung hoch verfügbar macht.

S

Scalable Coherent Interface (SCI) Hochgeschwindigkeits-Verbindungshardware, die als Cluster-Interconnect eingesetzt wird.

Scalable-Dienst Ein Datendienst, der so implementiert ist, dass er auf mehreren Knoten gleichzeitig ausgeführt werden kann.

Scalable-Ressource Eine Ressource, die auf mehreren Knoten läuft (eine Instanz auf jedem Knoten) und den Cluster-Interconnect verwendet, um den Remote-Clients den Eindruck eines einzigen Dienstes zu vermitteln.

Sekundärhostname Der Name eines Knotens, der für den Zugriff auf ein sekundäres öffentliches Netzwerk verwendet wird. Siehe auch “Primärhostname”.

Sekundärknoten Ein Cluster-Mitglied, das als Master für Platten- und Ressourcengruppen verfügbar ist, wenn der Primärknoten versagt. Siehe auch “Primärknoten”.

Sicherungsgruppe Siehe “IP Network Multipathing-Gruppe”.

Solaris Volume Manager Ein vom SunPlex-System verwendeter Datenträger-Manager. Siehe auch “Datenträger-Manager”.

Spare-Knoten Ein Cluster-Knoten, der bei einem Failover zum Sekundärknoten gemacht werden kann. Siehe auch “Sekundärknoten”.

Split Brain Ein Zustand, in dem ein Cluster in mehrere Partitionen zerfällt und jede Partition entsteht, ohne die Existenz irgend einer anderen Partition wahrzunehmen.

Standard-Master Das Cluster-Mitglied, auf dem standardmäßig ein Failover-Ressourcentyp online gebracht wird.

Sun Cluster (Software) Der Softwareanteil des SunPlex-Systems. (Siehe SunPlex.)

SunPlex Das integrierte Hardware- und Sun Cluster-Softwaresystem zum Erstellen hoch verfügbarer und Scalable-Dienste.

Switchback Siehe “Failback”.

Switchover Der planmäßige Wechsel einer Ressourcen- oder Gerätegruppe von einem Master (Knoten) in einem Cluster zu einem anderen Master (oder mehreren Mastern, wenn die Ressourcengruppen für mehrere

Primärknoten konfiguriert sind). Ein Switchover wird von einem Verwalter mit dem `scswitch(1M)`-Befehl eingeleitet.

System Service Processor (SSP) Ein Gerät, das sich in Enterprise 10000-Konfigurationen außerhalb des Clusters befindet und speziell für die Kommunikation mit Cluster-Mitgliedern eingesetzt wird.

T

Takeover Siehe “Failover”.

Terminal-Konzentrator Ein Gerät, das sich in anderen als Enterprise 10000-Konfigurationen außerhalb des Clusters befindet und speziell für die Kommunikation mit Cluster-Mitgliedern eingesetzt wird.

V

VERITAS Volume Manager Ein vom SunPlex-System verwendeter Datenträger-Manager. Siehe auch “Datenträger-Manager”.

Verwaltungskonsole Eine Workstation, auf der Cluster-Verwaltungssoftware ausgeführt wird.

Index

D

E

F

G

H

Q

R

S

T

U

V

W

Z