



Sun Cluster para el sistema operativo Solaris: Visión general

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

Referencia: 817-6910
Mayo 2004, Revisión A

Copyright 2004 Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 U.S.A. Reservados todos los derechos.

Este producto o documento está protegido por la ley de copyright y se distribuye bajo licencias que restringen su uso, copia, distribución y descompilación. No se puede reproducir parte alguna de este producto o documento en ninguna forma ni por cualquier medio sin la autorización previa por escrito de Sun y sus licenciadores, si los hubiera. El software de otras empresas, incluida la tecnología de los tipos de letra, está protegido por la ley de copyright y con licencia de los distribuidores de Sun.

Determinadas partes del producto pueden derivarse de Berkeley BSD Systems, con licencia de la Universidad de California. UNIX es una marca registrada en los EE.UU. y otros países, bajo licencia exclusiva de X/Open Company, Ltd.

Sun, Sun Microsystems, el logotipo de Sun, docs.sun.com, AnswerBook, AnswerBook2 y Solaris son marcas comerciales, marcas comerciales registradas o marcas de servicio de Sun Microsystems, Inc. en los EE.UU. y en otros países. Todas las marcas registradas SPARC se usan bajo licencia y son marcas comerciales o marcas registradas de SPARC International, Inc. en los EE.UU. y en otros países. Los productos con la marca comercial SPARC se basan en una arquitectura desarrollada por Sun Microsystems, Inc.

OPEN LOOK y la Interfaz gráfica de usuario (Graphical User Interface) de Sun™ fueron desarrollados por Sun Microsystems, Inc. para sus usuarios y licenciarios. Sun reconoce los esfuerzos pioneros de Xerox en la investigación y desarrollo del concepto de interfaces gráficas o visuales de usuario para la industria de la computación. Sun mantiene una licencia no exclusiva de Xerox para la interfaz gráfica de usuario de Xerox, que también cubre a los licenciarios de Sun que implementen GUI de OPEN LOOK y que por otra parte cumplan con los acuerdos de licencia por escrito de Sun.

Derechos del gobierno de los EE.UU. – Software comercial. Los usuarios del gobierno están sujetos a los términos de la licencia estándar de Sun Microsystems, Inc. y a las provisiones aplicables de las FAR y sus suplementos.

LA DOCUMENTACIÓN SE PROPORCIONA “TAL CUAL” Y SE RECHAZA CUALQUIER CONDICIÓN, REPRESENTACIÓN O GARANTÍA EXPLÍCITA O IMPLÍCITA, INCLUIDA CUALQUIER GARANTÍA IMPLÍCITA DE COMERCIABILIDAD, ADECUACIÓN A UN PROPÓSITO CONCRETO O NO INFRINGIMIENTO, EXCEPTO EN LO QUE DICHO RECHAZO ESTÉ INVALIDADO LEGALMENTE.

Copyright 2004 Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 U.S.A. Tous droits réservés.

Ce produit ou document est protégé par un copyright et distribué avec des licences qui en restreignent l'utilisation, la copie, la distribution, et la décompilation. Aucune partie de ce produit ou document ne peut être reproduite sous aucune forme, par quelque moyen que ce soit, sans l'autorisation préalable et écrite de Sun et de ses bailleurs de licence, s'il y en a. Le logiciel détenu par des tiers, et qui comprend la technologie relative aux polices de caractères, est protégé par un copyright et licencié par des fournisseurs de Sun.

Des parties de ce produit pourront être dérivées du système Berkeley BSD licenciés par l'Université de Californie. UNIX est une marque déposée aux Etats-Unis et dans d'autres pays et licenciée exclusivement par X/Open Company, Ltd.

Sun, Sun Microsystems, le logo Sun, docs.sun.com, AnswerBook, AnswerBook2, et Solaris sont des marques de fabrique ou des marques déposées, ou marques de service, de Sun Microsystems, Inc. aux Etats-Unis et dans d'autres pays. Toutes les marques SPARC sont utilisées sous licence et sont des marques de fabrique ou des marques déposées de SPARC International, Inc. aux Etats-Unis et dans d'autres pays. Les produits portant les marques SPARC sont basés sur une architecture développée par Sun Microsystems, Inc.

L'interface d'utilisation graphique OPEN LOOK et Sun™ a été développée par Sun Microsystems, Inc. pour ses utilisateurs et licenciés. Sun reconnaît les efforts de pionniers de Xerox pour la recherche et le développement du concept des interfaces d'utilisation visuelle ou graphique pour l'industrie de l'informatique. Sun détient une licence non exclusive de Xerox sur l'interface d'utilisation graphique Xerox, cette licence couvrant également les licenciés de Sun qui mettent en place l'interface d'utilisation graphique OPEN LOOK et qui en outre se conforment aux licences écrites de Sun.

CETTE PUBLICATION EST FOURNIE “EN L'ETAT” ET AUCUNE GARANTIE, EXPRESSE OU IMPLICITE, N'EST ACCORDEE, Y COMPRIS DES GARANTIES CONCERNANT LA VALEUR MARCHANDE, L'APTITUDE DE LA PUBLICATION A REpondre A UNE UTILISATION PARTICULIERE, OU LE FAIT QU'ELLE NE SOIT PAS CONTREFAISANTE DE PRODUIT DE TIERS. CE DENI DE GARANTIE NE S'APPLIQUERAIT PAS, DANS LA MESURE OU IL SERAIT TENU JURIDIQUEMENT NUL ET NON AVENU.



040511 @8606



Contenido

Prefacio 5

1 Introducción a Sun Cluster 9

- Alta disponibilidad real de las aplicaciones gracias a Sun Cluster 9
 - Gestión de disponibilidad 10
 - Recuperación de fallos, servicios escalables y aplicaciones paralelas 11
 - Ruta múltiple de red IP 11
 - Gestión del almacenamiento 12
- Supervisión de fallos 13
- Herramientas de configuración y de administración 14
 - SunPlex Manager 14
 - Interfaz de línea de órdenes 15
 - Sun Management Center 15
 - Control de acceso basado en el rol 15

2 Conceptos claves de Sun Cluster 17

- Nodos del clúster 17
- Interconexión del clúster 18
- Pertenencia al clúster 18
- Depósito de configuración del clúster 19
- Supervisores de errores 20
 - Supervisión de los servicios de datos 20
 - Supervisión de las rutas de los discos 20
 - Supervisión de ruta múltiple de red IP 20
- Dispositivos del quórum 21

	Integridad de los datos	21
	Aislamiento de fallos	22
	Mecanismo de recuperación rápida para el aislamiento de fallos	23
	Dispositivos	23
	Dispositivos globales	23
	Dispositivos locales	24
	Grupos de dispositivos de discos	25
	Servicios de datos	25
	Tipos de recursos	26
	Recursos	26
	Grupos de recursos	26
	Tipos de servicios de datos	26
3	Arquitectura de Sun Cluster	29
	Entorno de hardware de Sun Cluster	29
	Entorno del software de Sun Cluster	30
	Supervisor de pertenencia al clúster (CMM)	31
	Depósito de configuración del clúster (CCR)	32
	Sistemas de archivos del clúster	32
	Servicios de datos escalables	33
	Normas de equilibrio de cargas	34
	Almacenamiento en discos multisistema	35
	Interconexión del clúster	35
	Grupos de ruta múltiple de red IP	37
	Interfaces de red pública	37
	Índice	39

Prefacio

Sun™ Cluster para el sistema operativo Solaris: Visión general presenta el producto Sun Cluster, cuál es la finalidad de éste y qué medios usa para conseguir este fin; se explican también sus conceptos claves. La información de este documento permite familiarizarse con las funciones y las características de Sun Cluster.

Documentación relacionada

Puede encontrar información sobre temas referentes a Sun Cluster en la documentación enumerada en la tabla siguiente. Toda la documentación de Sun Cluster está disponible en <http://docs.sun.com>.

Tema	Documentación
Conceptos	<i>Sun Cluster Concepts Guide for Solaris OS</i>
Visión general	<i>Sun Cluster Overview for Solaris OS</i>
Glosario	<i>Sun Java Enterprise System 2003Q4 Glossary</i>
Administración del hardware	<i>Sun Cluster 3.x Hardware Administration Manual for Solaris OS</i> Guías individuales de administración del hardware
Instalación del software	<i>Sun Cluster Software Installation Guide for Solaris OS</i>
Administración de los servicios de datos	<i>Sun Cluster Data Services Planning and Administration Guide for Solaris OS</i> Guías individuales de los servicios de datos

Tema	Documentación
Desarrollo de los servicios de datos	<i>Sun Cluster Data Services Developer's Guide for Solaris OS</i>
Administración del sistema	<i>Sun Cluster System Administration Guide for Solaris OS</i>
Mensajes de error	<i>Sun Cluster Error Messages Guide for Solaris OS</i>
Referencias de órdenes y funciones	<i>Sun Cluster Reference Manual for Solaris OS</i>

Si desea una lista completa de la documentación de Sun Cluster, consulte las notas sobre la versión de Sun Cluster en <http://docs.sun.com>.

Acceso a la documentación de Sun en línea

La sede web docs.sun.comSM permite acceder a la documentación técnica de Sun en línea. Puede explorar el archivo docs.sun.com, buscar el título de un manual o un tema específicos. El URL es <http://docs.sun.com>.

Solicitud de documentación de Sun

Sun Microsystems ofrece una seleccionada documentación impresa sobre el producto. Si desea saber la lista completa de documentos y cómo solicitarlos, consulte "Adquirir documentación impresa" en <http://docs.sun.com>.

Obtención de ayuda

Si tiene problemas durante la instalación o el uso del sistema Sun Cluster, póngase en contacto con su proveedor de servicios y déle la información siguiente:

- Su nombre y dirección de correo electrónico (si estuviera disponible)
- El nombre, dirección y número de teléfono de su empresa

- Los modelos y números de serie de sus sistemas
- El número de versión del sistema operativo; por ejemplo Solaris 9
- El número de versión del software Sun Cluster (por ejemplo, 3.1 4/04)

Use las órdenes siguientes para reunir la información sobre todos los nodos del sistema que necesitará el proveedor de servicio.

Orden	Función
<code>prtconf -v</code>	Muestra el tamaño de la memoria del sistema y ofrece información sobre los dispositivos periféricos
<code>psrinfo -v</code>	Muestra información acerca de los procesadores
<code>showrev -p</code>	Indica las modificaciones instaladas
<code>prtdiag -v</code>	Muestra información de diagnóstico del sistema
<code>scinstall -pv</code>	Muestra información sobre la versión y el paquete de Sun Cluster.
<code>scstat</code>	Ofrece una instantánea del estado del clúster.
<code>scconf -p</code>	Muestra información de configuración sobre el clúster
<code>scrgadm -p</code>	Muestra información sobre los recursos instalados, los grupos y tipos de recursos

Tenga también a punto el contenido del archivo `/var/adm/messages`.

Convenciones tipográficas

La tabla siguiente describe los cambios tipográficos utilizados en este manual.

TABLA P-1 Convenciones tipográficas

Tipo de letra o símbolo	Significado	Ejemplo
AaBbCc123	Los nombres de las órdenes, los archivos y los directorios, así como la salida por la pantalla del computador	Edite el archivo <code>.login</code> . Utilice <code>ls -a</code> para mostrar una lista de todos los archivos. <code>nombre_sistema%</code> tiene correo.

TABLA P-1 Convenciones tipográficas (Continuación)

Tipo de letra o símbolo	Significado	Ejemplo
AaBbCc123	Lo que usted escribe, contrastado con la salida por la pantalla del computador	nombre_sistema% su Password:
<i>AaBbCc123</i>	Plantilla de línea de órdenes: sustituir por un valor o nombre real	Para suprimir un archivo, escriba rm nombre_archivo.
<i>AaBbCc123</i>	Títulos de manuales, términos o expresiones nuevas, palabras que deben destacarse.	Véase el capítulo 6 de la <i>Guía del usuario</i> Se denominan opciones de <i>clase</i> . Para hacer esto debe ser el usuario <i>root</i> .

Indicadores de los shells en los ejemplos de órdenes

La tabla siguiente muestra los indicadores predeterminados del sistema y de superusuario para los shells Bourne, Korn y C.

TABLA P-2 Indicadores de los shells

Shell	Indicador
Indicador del shell C	nombre_sistema%
Indicador de superusuario en el shell C	nombre_sistema#
Indicador de los shells Bourne y Korn	\$
Indicador de superusuario en los shell Bourne y Korn	#

Introducción a Sun Cluster

El sistema SunPlex es una solución integrada de hardware y software de Sun Cluster que se utiliza para crear servicios de alta disponibilidad y escalabilidad. Este capítulo proporciona una visión general muy aclaratoria sobre las funciones de Sun Cluster.

Este capítulo se divide en los siguientes apartados:

- “Alta disponibilidad real de las aplicaciones gracias a Sun Cluster” en la página 9
- “Supervisión de fallos” en la página 13
- “Herramientas de configuración y de administración” en la página 14

Alta disponibilidad real de las aplicaciones gracias a Sun Cluster

Un clúster son dos sistemas (o nodos) o más que trabajan conjuntamente como un único sistema permanentemente disponible, con el fin de proporcionar aplicaciones, recursos de sistemas y datos a los usuarios. Cada nodo de un clúster es un sistema autónomo completamente operativo. No obstante, en un entorno de clústers, los nodos están conectados mediante una interconexión y funcionan conjuntamente como una única entidad con el fin de proporcionar una disponibilidad y un rendimiento mejores.

La alta disponibilidad de los clústers proporciona un acceso casi continuo a los datos y aplicaciones, manteniendo al clúster en funcionamiento, aun en el caso de fallos que normalmente bloquearían un único servidor. Un único fallo en el hardware, en el software o en la red no harán que el clúster falle. Por contra, los sistemas de hardware tolerante a fallos proporcionan un acceso continuado a datos y aplicaciones, pero a un coste más elevado debido al uso de hardware especializado. Los sistemas tolerantes a fallos, normalmente, no previenen los fallos de software.

Cada sistema Sun Cluster es un conjunto de nodos perfectamente acoplados que proporcionan una única visión de la administración de los servicios de red y aplicaciones. El sistema Sun Cluster consigue una alta disponibilidad en la combinación del hardware y del software siguientes:

- Los sistemas de discos redundantes proporcionan una mayor capacidad de almacenamiento. Estos sistemas de discos generalmente se duplican para permitir un funcionamiento ininterrumpido en el caso de que falle un disco o un subsistema. Las conexiones redundantes con los sistemas de discos aseguran que los datos no queden aislados en el caso de que falle un servidor, un controlador o un cable. Una interconexión de alta velocidad entre los nodos proporciona acceso a los recursos. Todos los nodos del clúster también están conectados con una red pública, lo que permite a los clientes de varias redes acceder al clúster.
- Los componentes redundantes intercambiables en marcha, como las fuentes de alimentación y los sistemas de refrigeración, mejoran la disponibilidad, puesto que permiten a los sistemas continuar funcionando tras un error del hardware. Los componentes intercambiables en marcha proporcionan la posibilidad de añadir o suprimir componentes de hardware en un sistema en funcionamiento, sin bloquearlo.
- La estructura de alta disponibilidad del software Sun Cluster detecta rápidamente un error en el nodo y migra la aplicación o servicio a otro nodo que se ejecute en un entorno idéntico. Siempre habrá alguna aplicación disponible. Las aplicaciones no se ven afectadas por un nodo bloqueado y están totalmente disponibles durante el proceso de recuperación. Además, las aplicaciones del nodo fallido vuelven a estar disponibles en cuanto se recuperan, de esta manera no tienen que esperar a que todas las demás aplicaciones terminen de recuperarse.

Gestión de disponibilidad

Una aplicación está realmente disponible si sobrevive a cualquier fallo (de software o hardware) que se produzca en el sistema. Se excluyen los fallos provocados por errores o por el deterioro de datos en la propia aplicación. La información siguiente se aplica a las aplicaciones altamente disponibles:

- La recuperación es transparente a partir de las aplicaciones que utilizan un recurso.
- El acceso a los recursos está totalmente garantizado durante el error de un nodo.
- Las aplicaciones no pueden detectar si el nodo de alojamiento se ha movido a otro.
- El error de un único nodo es completamente transparente para los programas en los nodos restantes que utilicen los archivos, los dispositivos y los volúmenes de discos acoplados a este nodo.

Recuperación de fallos, servicios escalables y aplicaciones paralelas

La recuperación de fallos, los servicios escalables y las aplicaciones paralelas permiten conseguir que las aplicaciones estén realmente disponibles y mejoran el rendimiento de una aplicación en un clúster.

Un servicio de recuperación de fallos proporciona una alta disponibilidad gracias a la redundancia. Cuando se produce un fallo, se puede configurar una aplicación que se esté ejecutando para reiniciarla en el mismo nodo o para moverla a otro nodo del clúster, sin la intervención del usuario.

Si desea aumentar el rendimiento, un servicio escalable integra a los diferentes nodos de un clúster para ejecutar una aplicación al mismo tiempo. En una configuración escalable, todos los nodos del clúster pueden proporcionar datos y procesar peticiones de los clientes.

Las bases de datos paralelas permiten que varias instancias del servidor de la base de datos:

- Participen en el clúster
- Manejen simultáneamente varias consultas en la misma base de datos
- Proporcionen la posibilidad de efectuar consultas paralelas en las grandes consultas

Si desea obtener más información sobre los servicios escalables y de recuperación de fallos y las aplicaciones paralelas, consulte “Tipos de servicios de datos” en la página 26.

Ruta múltiple de red IP

Los clientes hacen peticiones de datos al clúster a través de la red pública. Cada nodo del clúster está conectado como mínimo a una red pública a través de uno o varios adaptadores.

Ruta múltiple de red IP permite a un servidor disponer de varios puertos de red conectados con la misma subred. En primer lugar, el software Ruta múltiple de red IP proporciona capacidad de recuperación a partir de fallos en los adaptadores de red mediante la detección del fallo o la reparación de un adaptador de red, a continuación, conmuta simultáneamente la dirección de la red con y desde un adaptador alternativo. Si hay operativo más de un adaptador de red, Ruta múltiple de red IP aumenta el caudal de datos distribuyendo los paquetes de salida entre los adaptadores.

Gestión del almacenamiento

El almacenamiento multisistema consigue que los discos estén realmente disponibles mediante la conexión de los discos con varios nodos los cuales permiten que haya diferentes rutas para acceder a los datos; si una de éstas falla habrá otra disponible para ocupar su lugar.

Los discos multisistema permiten los procesos de clústers siguientes:

- Tolerar los fallos en un único nodo.
- Centralizar los datos de las aplicaciones, los archivos binarios de las aplicaciones y los archivos de configuración.
- Conseguir la protección frente a los fallos en los nodos. Si las peticiones de los clientes están accediendo a datos a través de un nodo que falla, se desvían para usar otro nodo que tenga una conexión directa con los mismos discos.
- Proporcionar acceso globalmente a través de un nodo principal que “controle” los discos o mediante acceso directo simultáneo a través de las rutas locales.

Compatibilidad con la gestión de volúmenes

Un gestor de volúmenes permite gestionar un gran número de discos y los datos que éstos contienen. Los gestores de volúmenes pueden aumentar la capacidad de almacenamiento y la disponibilidad de los datos mediante estas funciones:

- Concatenación y reparto en bandas de las unidades de discos
- Duplicación de discos
- Sustitución de unidades de discos en marcha
- Gestión de los errores de los discos y la sustitución de éstos

Los sistemas Sun Cluster admiten los gestores de volúmenes siguientes:

- Solaris Volume Manager
- VERITAS Volume Manager

Sun StorEdge Traffic Manager

El software Sun StorEdge Traffic Manager está completamente integrado a partir de la estructura E/S central de Sistema operativo Solaris 8; permite representar y gestionar de manera más efectiva recursos, accesibles a través de varias interfaces de controladores E/S en una única instancia del entorno operativo Solaris. La arquitectura de Sun StorEdge Traffic Manager permite:

- Protección frente a las interrupciones de E/S debidas a fallos en los controladores de E/S
- Conmutación automática con un controlador alternativo tras un fallo del controlador de E/S

- Rendimiento mejorado de E/S, equilibrando para ello la carga en varios canales de E/S

Admisión de una matriz redundante de hardware de discos independientes

Los sistemas Sun Cluster admiten el uso de una Matriz redundante de hardware de discos independientes (RAID) y de software RAID basado en el sistema. El hardware RAID usa la redundancia de hardware del sistema de almacenamiento o de la matriz de almacenamiento para asegurarse de que los fallos del hardware independiente no tengan consecuencias sobre la disponibilidad de los datos. Si efectúa una duplicación en matrices de discos separadas, el software RAID basado en el sistema garantiza que los fallos del hardware independiente no repercutan en la disponibilidad de los datos si una matriz de almacenamiento completa queda fuera de línea. Aunque pueda usar el hardware RAID y el software RAID basado en sistemas, sólo necesita una solución RAID para mantener un alto grado de disponibilidad de los datos.

Admisión de sistemas de archivos

Puesto que una de las propiedades inherentes a los sistemas de un clúster es la de compartir recursos, éste necesita un sistema de archivos que controle que la necesidad de archivos se comparta de manera equitativa. El sistema de archivos de Sun Cluster permite a los usuarios o a las aplicaciones acceder a un archivo de un nodo del clúster, mediante las API estándar de UNIX remotas o locales. Si una aplicación se traslada de un nodo a otro, no se necesita ningún cambio en la aplicación para acceder a los mismos archivos. No se necesitan cambios en las aplicaciones para utilizar el sistema de archivos del clúster.

Supervisión de fallos

El sistema Sun Cluster consigue que la ruta entre los usuarios y los datos esté completamente disponible mediante los discos multisistema, la ruta múltiple y un sistema de archivos global. El sistema Sun Cluster supervisa los fallos de:

- Aplicaciones: la mayoría de los servicios de datos de Sun Cluster proporcionan un supervisor de fallos que regularmente comprueba los servicios de datos, con el fin de determinar su estado. Un supervisor de fallos comprueba que los daemons de la aplicación se estén ejecutando y que los clientes estén atendidos. Según la información que devuelvan los análisis, se puede iniciar una acción predefinida como el reinicio de los daemons o una recuperación de fallos.

- Rutas de discos: Sun Cluster admite la supervisión de rutas de discos (DPM) que mejora la fiabilidad general de la recuperación de fallos y la conmutación informando del fallo de una ruta de discos secundaria.
- Protocolo de Internet (IP) de ruta múltiple: el software de ruta múltiple de la red IP de Solaris de los sistemas Sun Cluster proporciona el mecanismo básico para la supervisión de adaptadores de red pública. La ruta múltiple IP también permite la recuperación de fallos de direcciones IP desde un adaptador a otro si se detecta un fallo.

Herramientas de configuración y de administración

El sistema Sun Cluster se puede instalar, configurar y administrar mediante las interfaces SunPlex Manager o de la línea de órdenes (CLI).

El sistema Sun Cluster también dispone de un módulo que se ejecuta como parte del software Sun Management Center que proporciona una interfaz gráfica de usuario (GUI) para ciertas tareas del clúster.

SunPlex Manager

SunPlex Manager es una herramienta basada en el navegador para administrar sistemas Sun Cluster. El sistema SunPlex Manager permite a los administradores efectuar la supervisión y la gestión del sistema, la instalación del software y la configuración del sistema.

El software SunPlex Manager tiene las funciones siguientes.

- Mecanismos de autorización y de seguridad incorporada
- Compatibilidad con Secure Sockets Layer (SSL, capa de zócalos protegidos)
- Control de acceso basado en el rol (RBAC)
- Módulo de autenticación conectable (PAM)
- Recursos de administración de grupos de ruta múltiple de la red IP y NAFO
- Administración de dispositivos del quórum, transportes, dispositivos de almacenamiento compartido y grupos de recursos
- Avanzada comprobación de errores y detección automática de interconexiones privadas

Interfaz de línea de órdenes

La interfaz de línea de órdenes de Sun Cluster es un conjunto de utilidades que permiten instalar y administrar los sistemas Sun Cluster, así como administrar la parte del gestor de volúmenes del software Sun Cluster.

La interfaz de línea de órdenes de Sun Cluster permite efectuar las siguientes tareas de administración de SunPlex:

- Validar una configuración de Sun Cluster
- Instalar y configurar Sun Cluster
- Actualizar una configuración de Sun Cluster
- Gestionar el registro de los tipos de recursos, crear grupos de recursos y activar recursos dentro de un grupo de recursos.
- Cambiar el dominio del nodo, el estado de los grupos de recursos y los grupos de dispositivos de discos
- Controlar el acceso mediante el Control de acceso basado en el rol (RBAC)
- Apagar todo el clúster

Sun Management Center

El sistema Sun Cluster también cuenta con un módulo que se ejecuta como parte del software Sun Management Center. El clúster se basa en éste para realizar las funciones de administración y supervisión. Las siguientes tareas se pueden efectuar a través de una interfaz de línea de órdenes (CLI) o una interfaz gráfica de usuario (GUI):

- Configurar un sistema remoto
- Supervisar el rendimiento
- Detectar y aislar los errores de hardware y de software

También es posible utilizar el software Sun Management Center como interfaz para gestionar la reconfiguración dinámica en los servidores de Sun Cluster. Ésta incluye la creación de dominios, la conexión de tarjetas dinámicas y la desconexión dinámica.

Control de acceso basado en el rol

En los sistemas UNIX convencionales el usuario root, también conocido como superusuario, es omnipotente: puede leer y guardar datos en cualquier archivo, ejecutar todos los programas y enviar señales de finalización a cualquier proceso. El control de acceso basado en el rol (RBAC) de Solaris es una alternativa al modelo de superusuario “todo o nada”; usa el principio de seguridad del privilegio mínimo, según el cual ningún usuario debe recibir más privilegios que los necesarios para llevar a cabo su trabajo.

RBAC permite a una organización distribuir y empaquetar los permisos o derechos del superusuario en funciones o cuentas especiales de usuario que se pueden asignar a individuos determinados, de esta manera se pueden establecer diferentes normas de seguridad. Las cuentas se pueden configurar para administradores con funciones especiales en áreas como la seguridad, el trabajo en red, los cortafuegos, las copias de seguridad y el funcionamiento del sistema.

Conceptos claves de Sun Cluster

Este capítulo explica los conceptos claves relacionados con los componentes de hardware y software del sistema Sun Cluster que es necesario comprender para comenzar a trabajar con los sistemas Sun Cluster.

Este capítulo se divide en los siguientes apartados:

- “Nodos del clúster” en la página 17
- “Interconexión del clúster” en la página 18
- “Pertenencia al clúster” en la página 18
- “Depósito de configuración del clúster” en la página 19
- “Supervisores de errores” en la página 20
- “Dispositivos del quórum” en la página 21
- “Dispositivos” en la página 23
- “Servicios de datos” en la página 25

Nodos del clúster

Un nodo del clúster es una máquina que ejecuta los softwares Solaris y Sun Cluster. Éste permite tener de dos a ocho nodos por clúster.

Los nodos del clúster se acoplan generalmente a uno o más discos; aquellos que no lo están usan el sistema de archivos del clúster para acceder a los discos multisistema. Los nodos en configuraciones de bases de datos paralelas comparten acceso simultáneo a algunos de los discos o a todos ellos.

Todos los nodos del clúster reciben información de cuándo un nodo se une al clúster o deja éste, conocen también los recursos que se están ejecutando tanto localmente como en los otros nodos del clúster.

Los nodos del mismo clúster deben tener capacidades de procesamiento, memoria y E/S similares para permitir que la recuperación de fallos se produzca sin que haya una degradación importante en el rendimiento. Debido a la posibilidad de la recuperación de un fallo, cada nodo debe tener capacidad suficiente para cumplir con los acuerdos de los niveles de servicios si un nodo falla.

Interconexión del clúster

La interconexión del clúster es la configuración física de dispositivos que se utiliza para la transferencia de comunicaciones privadas de clústers y comunicaciones de servicios de datos entre los nodos de los clústers.

Las interconexiones redundantes permiten que continúe el funcionamiento en las interconexiones que queden, mientras los administradores de los sistemas aíslan los fallos y restablecen la comunicación. Sun Cluster detecta, repara y reinicia automáticamente la comunicación en una interconexión restablecida.

Si desea más información, consulte “Interconexión del clúster” en la página 35.

Pertenencia al clúster

El Supervisor de pertenencia al clúster (CMM) es un conjunto distribuido de agentes que intercambian mensajes en una interconexión de clústers con el fin de terminar las tareas siguientes:

- Forzar una vista de miembros uniforme en todos los nodos (quórum)
- Controlar una reconfiguración sincronizada en respuesta a los cambios en la pertenencia
- Gestionar las particiones del clúster
- Asegurar una total conectividad entre todos los miembros de los clústers, dejando los nodos dañados fuera del clúster hasta que se reparen

La función principal del CMM es establecer la pertenencia al clúster, lo cual requiere un acuerdo en todo el clúster en el conjunto de nodos que forman parte del clúster en cualquier momento. El CMM detecta cambios importantes en el estado de los clústers en cada nodo, como, por ejemplo, una pérdida de comunicación entre uno o más nodos; confía en el módulo del núcleo de transporte para generar pulsos en el medio de transporte con otros nodos del clúster y, si no detecta un pulso de un nodo en un período concreto, considera que el nodo ha fallado e inicia una reconfiguración del clúster para renegociar la pertenencia a éste.

Para determinar ésta y para asegurar la integridad de los datos, el CMM efectúa las tareas siguientes:

- Registrar los cambios en la pertenencia al clúster, como la incorporación o el cese de un nodo en el clúster
- Asegurarse de que el nodo dañado deje de estar en el clúster
- Asegurarse de que un nodo dañado permanezca inactivo hasta que se repare
- Evitar que el clúster se particione en subgrupos de nodos

Consulte “Integridad de los datos” en la página 21 si desea más información sobre cómo se protege el clúster de particionarse en varios clústers independientes.

Depósito de configuración del clúster

El Depósito de configuración del clúster (CCR) es una base de datos privada, distribuida en todo el clúster para almacenar la información que pertenece a la configuración y al estado del clúster. Con el fin de evitar el deterioro de los datos de la configuración, cada nodo debe tener información sobre el estado actual de los recursos de los clústers. El CCR se asegura de que todos los nodos tengan una visión coherente del clúster, para ello se actualiza cuando se produce un error o una recuperación de un fallo o cuando el estado general del clúster cambia.

Las estructuras del CCR contienen los tipos siguientes de información:

- Nombres de los nodos y de los clústers
- Configuración para el transporte de los clústers
- Los nombres de los conjuntos de discos Solaris Volume Manager o los grupos de discos VERITAS
- Una lista de los nodos que puede controlar cada grupo de discos
- Valores operativos de parámetros para los servicios de datos
- Rutas a los métodos de rellamada de los servicios de datos
- Configuración de dispositivos DID
- Estado actual del clúster

Supervisores de errores

El sistema Sun Cluster permite que todos los componentes de la “ruta” entre los usuarios y los datos queden totalmente disponibles, al supervisar las aplicaciones, el sistema de archivos y las interfaces de la red.

Sun Cluster detecta rápidamente un fallo en el nodo y crea un servidor equivalente para los recursos en el nodo que ha fallado. Sun Cluster se asegura de que los recursos que no resulten afectados por el nodo incorrecto estén siempre disponibles durante la recuperación y que los recursos del nodo defectuoso queden disponibles tan pronto como se recuperen.

Supervisión de los servicios de datos

Cada servicio de datos de Sun Cluster proporciona un supervisor de fallos que explora periódicamente el servicio de datos para determinar su buen estado. Un supervisor de errores comprueba que el daemon de la aplicación o los daemons se estén ejecutando y que los clientes reciban servicio. Basándose en la información proporcionada por los análisis, se pueden iniciar acciones predefinidas como el reinicio de los daemons o la activación de la recuperación de un fallo.

Supervisión de las rutas de los discos

Sun Cluster admite la supervisión de las rutas de los discos (DPM, disk-path monitoring) que mejora la fiabilidad general en la recuperación de fallos y en la conmutación, al informar sobre el fallo de una ruta de discos secundaria. Para supervisar las rutas de los discos hay dos métodos disponibles. El primero, usar la orden `scdpm` que permite supervisar, dejar de supervisar o mostrar el estado de las rutas del disco del clúster. Consulte la página de comando `man scdpm(1M)` para obtener más información sobre las opciones de la línea de órdenes.

El segundo método es utilizar la interfaz gráfica del usuario (GUI) de SunPlex Manager que proporciona una visión topológica de las rutas de discos supervisadas. La vista se actualiza cada 10 minutos para incluir información sobre el número de pings que han fallado.

Supervisión de ruta múltiple de red IP

Todos los nodos del clúster tienen su propia configuración de Ruta múltiple de red IP, que puede ser distinta de la configuración de otros nodos del clúster. Ruta múltiple de red IP supervisa los siguientes errores en la comunicación a través de la red:

- La ruta de transmisión y recepción del adaptador de red ha detenido la transmisión de paquetes.
- El acoplamiento de los adaptadores de la red con el enlace está desactivado.
- El puerto del interruptor no transmite ni recibe paquetes.
- La interfaz física de un grupo no está presente en el arranque del sistema.

Dispositivos del quórum

Un dispositivo del quórum es un disco compartido por dos nodos o más que aporta votos que se utilizan para establecer un quórum con el fin de que se ejecute el clúster, ya que éste sólo puede funcionar cuando está disponible un quórum de votos. El dispositivo del quórum se usa cuando un clúster se particiona en distintos conjuntos de nodos, para establecer qué conjunto de nodos constituye el nuevo clúster.

Los nodos del clúster y los dispositivos del quórum votan para formar el quórum. De forma predeterminada los nodos del clúster aumentan en uno su recuento de votos del quórum cuando arrancan y se convierten en miembros del clúster. Los nodos pueden tener un recuento de votos de cero si el nodo se está instalando o cuando un administrador lo ha situado en estado de mantenimiento.

Los dispositivos del quórum adquieren votos según el número de conexiones de nodo que tenga el dispositivo. Cuando se configura un dispositivo del quórum, éste adquiere un recuento máximo de votos de $N-1$ donde N es el número de votos conectados al dispositivo del quórum. Por ejemplo, un dispositivo del quórum conectado a dos nodos con recuentos distintos de cero, tiene un recuento del quórum igual a uno (dos menos uno).

Integridad de los datos

El sistema Sun Cluster intenta evitar el deterioro de los datos y asegurar su integridad. Puesto que los nodos de los clústers comparten datos y recursos, un clúster nunca debe dividirse en particiones que estén activas al mismo tiempo. El CMM garantiza que sólo pueda haber un clúster operativo al mismo tiempo.

Pueden surgir dos tipos de problemas derivados de las particiones del disco: esquizofrenia y amnesia. La esquizofrenia se produce cuando la interconexión del clúster entre los nodos se pierde y el clúster se divide en subclústers, cada uno de los cuales cree que es la única partición. Un subclúster que desconoce la existencia de los otros puede provocar conflictos en los recursos compartidos como direcciones de red duplicadas y deterioro de datos.

La amnesia se produce si todos los nodos dejan el clúster en grupos residuales. Un ejemplo es un clúster de dos nodos A y B. Si el nodo A queda inactivo, los datos de la configuración del CCR se actualizan solamente en el nodo B, no en el A. Si el nodo B queda inactivo posteriormente y se rearranca el nodo A, éste se ejecutará con los antiguos contenidos del CCR. Este estado recibe el nombre de amnesia y puede llevar a ejecutar un clúster con información sobre la configuración del estado.

Para evitar la esquizofrenia y la amnesia se debe dar cada nodo un voto y obligar a que haya una mayoría de votos por clúster en funcionamiento. Una partición con la mayoría de votos tiene quórum y se le permite funcionar. Este mecanismo de voto por mayoría funciona bien si en el clúster hay más de dos nodos. En un clúster de dos nodos, la mayoría es dos. Si ese tipo del clúster resulta particionado, un voto externo permite que una partición obtenga el quórum. Este voto externo lo proporciona un dispositivo del quórum. Cualquier disco compartido entre los dos nodos puede ser un dispositivo del quórum.

La Tabla 2-1 describe cómo Sun Cluster utiliza el quórum para evitar la esquizofrenia y la amnesia.

TABLA 2-1 Quórum del clúster y problemas de esquizofrenia y amnesia

Tipo de partición	Solución del quórum
Esquizofrenia	Permite solamente la partición (subclúster) con una mayoría de votos que ejecutar como clúster (sólo puede existir una partición con dicha mayoría). Si el nodo pierde la votación y no reúne el quórum necesario, emite un aviso grave.
Amnesia	Garantiza que cuando se arranque un clúster, tenga al menos un nodo que era miembro de la propiedad del clúster más reciente (y por tanto disponga de los datos de configuración más recientes)

Aislamiento de fallos

Un problema fundamental de los clústers es un fallo que provoque en ellos una partición (denominada *esquizofrenia*). Cuando esto ocurre, no todos los nodos pueden comunicarse, por lo que algunos podrían intentar formar clústers individuales o subconjuntos que se “creerían” con permisos de acceso y de propiedad exclusivos respecto a los discos multisistema. Si varios nodos intentan guardar datos en los discos se puede producir un deterioro en los datos.

El aislamiento de fallos limita el acceso de los nodos a los discos multisistema, evitando que físicamente se pueda acceder a ellos. Cuando un nodo abandona el clúster (falla o se particiona), el aislamiento de fallos se asegura de que el nodo ya no pueda acceder a los discos. Sólo los nodos miembros actuales tendrán acceso a los discos, conservándose así la integridad de los datos.

El sistema Sun Cluster usa reservas de disco SCSI para implementar el aislamiento de fallos, gracias a las cuales, los nodos fallidos se “aislan” de los discos multisistema, evitando que accedan a estos discos.

Cuando un miembro del clúster detecta que otro nodo ya no se está comunicando a través de la interconexión del clúster, inicia un procedimiento de aislamiento de fallos para evitar que el nodo defectuoso acceda a los discos compartidos. En este proceso, el nodo excluido emite avisos graves y aparece un mensaje de “conflicto de reserva” en la consola.

Mecanismo de recuperación rápida para el aislamiento de fallos

El mecanismo de recuperación rápida emite avisos graves acerca de un nodo defectuoso, pero no evita que éste re arranque. Después del aviso grave, es posible que el nodo re arranque e intente reunificar el clúster.

Si un nodo pierde conectividad con los otros nodos del clúster y no es parte de una partición que pueda conseguir el quórum, otro nodo lo expulsa del clúster. Otro nodo que forma parte de la partición que puede conseguir reservas de plazas del quórum en los discos compartidos. El nodo que no tiene quórum emite después un aviso grave como resultado del mecanismo de recuperación rápida.



Dispositivos

El sistema de archivos global permite que todos los archivos del clúster sean accesibles por igual y que estén visibles para todos los nodos. De similar modo, Sun Cluster consigue que todos los dispositivos de un clúster sean accesibles y visibles en todo el clúster. Esto es, el subsistema de E/S permite el acceso a cualquier dispositivo del clúster, desde cualquier nodo, sin tener en cuenta dónde se acopla físicamente el dispositivo. Este acceso recibe el nombre de acceso global a dispositivos.

Dispositivos globales

Los sistemas Sun Cluster utilizan dispositivos globales con el fin de proporcionar un acceso realmente disponible en todo el clúster a cualquier dispositivo del clúster, desde cualquier nodo. Por lo general, si un nodo falla a la hora de proporcionar acceso a un dispositivo global, Sun Cluster conmuta a otra ruta al dispositivo y vuelve a dirigir el acceso a esa ruta. Esta redirección es fácil con los dispositivos globales, puesto que se utiliza el mismo nombre para el dispositivo, sin tener en cuenta la ruta.

El acceso a un dispositivo remoto se lleva a cabo del mismo modo que en un dispositivo local que utilice el mismo nombre. Asimismo, la API que se usa para acceder a un dispositivo global en un clúster es la misma que la que se utiliza para acceder de manera local a un dispositivo.

Los dispositivos globales de Sun Cluster pueden ser discos, CD-ROM y cintas. No obstante, los discos son los únicos dispositivos globales de puerto múltiple que se admiten. Ello significa que los CD-ROM y los dispositivos de cinta actualmente no son dispositivos de alta disponibilidad. Los discos locales de cada servidor tampoco son multipuerto, por lo tanto no son dispositivos de alta disponibilidad.

El clúster asigna ID exclusivos a cada disco, CD-ROM y dispositivo de cinta del clúster, lo que permite el acceso uniforme a todos los dispositivos desde cualquier nodo del clúster.

ID del dispositivo

El software Sun Cluster gestiona dispositivos globales a través de una estructura conocida como el controlador del ID del dispositivo (DID) que se utiliza para asignar automáticamente ID exclusivos a cada dispositivo del clúster, incluidos los discos multisistema, las unidades de cinta y los CD-ROM.

También es una pieza integral de la función de acceso global a los dispositivos del clúster. Asimismo, analiza todos los nodos del clúster y construye una lista de dispositivos exclusivos del disco. Este controlador también asigna a cada dispositivo un número menor y mayor exclusivos, coherentes en todos los nodos del clúster. El acceso a los dispositivos globales se realiza a través del DID exclusivo asignado por el controlador del DID en lugar de los DID tradicionales de Solaris.

De esta manera se asegura que cualquier aplicación que acceda a los discos, como Solaris Volume Manager o Sun Java System Directory Server, utilice una ruta coherente en todo el clúster. Esta coherencia es especialmente importante en los discos multisistema, puesto que los números locales menores y mayores de cada dispositivo pueden variar según el nodo y también pueden cambiar la convención de asignación de nombres del dispositivo Solaris.

Dispositivos locales

El software Sun Cluster gestiona también los dispositivos locales que son accesibles tan solo en un nodo que ejecute un servicio y que tenga una conexión física con el clúster. Es posible que los dispositivos locales tengan un mejor rendimiento que los globales, puesto que aquéllos no tienen que duplicar la información del estado en varios nodos simultáneamente. El fallo en el dominio del dispositivo suprime el acceso a éste a menos que varios nodos puedan compartir este dispositivo.

Grupos de dispositivos de discos

Los grupos de dispositivos de discos permiten a los grupos de discos del gestor de volúmenes convertirse en “globales”, gracias a la compatibilidad de ruta múltiple y de sistema múltiple en los discos subyacentes. Cada nodo del clúster conectado físicamente a los discos multisistema proporciona una ruta de acceso al grupo de dispositivos de disco.

En el sistema Sun Cluster, los discos de sistema múltiple pueden estar bajo el control del software Sun Cluster registrándose como grupos de dispositivos de discos. Este registro proporciona al sistema Sun Cluster información sobre qué nodos tienen una ruta a ciertos grupos de discos del gestor de volúmenes. El software de Sun Cluster crea un grupo de dispositivos básicos de disco para cada dispositivo de disco y de cinta del clúster. Estos grupos de dispositivos del clúster permanecen fuera de línea hasta que se accede a ellos como dispositivos globales montando un sistema de archivos global o accediendo a un archivo básico de bases de datos.

Servicios de datos

Un servicio de datos es la combinación de software y archivos de configuración que permite a una aplicación ejecutarse sin modificaciones en una configuración de Sun Cluster, ya que en esta circunstancia se ejecuta como recurso bajo el control del Gestor del grupo de recursos (RGM). Un servicio de datos permite configurar una aplicación como Sun Java System Web Server o la base de datos de Oracle para ejecutarse en un clúster en lugar de un único servidor.

El software de un servicio de datos proporciona implementaciones de los métodos de gestión de Sun Cluster que efectúan las operaciones siguientes en la aplicación:

- Iniciar la aplicación
- Parar la aplicación
- Supervisar los errores de la aplicación y recuperar después de estos fallos

Los archivos de configuración de un servicio de datos definen las propiedades del recurso que representa la aplicación en el gestor del grupo de recursos.

El RGM (gestor del grupo de recursos) controla la disposición de la recuperación de fallos y los servicios de datos escalables en el clúster y es responsable del inicio y de la parada de los servicios de datos en los nodos seleccionados del clúster, en respuesta a los cambios en la pertenencia al clúster; permite a las aplicaciones de servicios de datos utilizar la estructura del clúster.

El RGM controla los servicios de datos como recursos. Estas implementaciones las suministra Sun o las crea un desarrollador que utiliza una plantilla de servicios de datos genéricos, la API de la biblioteca de desarrollo de servicios de datos (DSDL API)

o la API de la gestión de recursos (RMAPI). El administrador del clúster crea y gestiona los recursos en contenedores que reciben el nombre de grupos de recursos. Las acciones del administrador y del RGM provocan que los recursos y los grupos de recursos estén alternativamente en línea y fuera de línea.

Tipos de recursos

Un tipo de recurso es un conjunto de propiedades que describen una aplicación al clúster y que contiene información sobre cómo se debe iniciar, detener y supervisar la aplicación en los nodos del clúster; incluye, también, las propiedades específicas de la aplicación que es necesario definir para usar ésta en el clúster. Los servicios de datos de Sun Cluster tienen varios tipos de recursos predefinidos. Por ejemplo, Sun Cluster HA para Oracle es el tipo de recurso `SUNW.oracle-server` y Sun Cluster HA for Apache es `SUNW.apache`.

Recursos

Un recurso es una instancia de un tipo de recurso que está definido en todo el clúster y que permite que varias instancias de una aplicación se instalen en éste. Cuando se inicializa un recurso, RGM asigna valores a las propiedades específicas de la aplicación y el recurso hereda las propiedades en el tipo de recurso.

Los servicios de datos usan varios tipos de recursos. Las aplicaciones como Apache Web Server o Sun Java System Web Server usan direcciones de red (nombres de sistema lógicos y direcciones compartidas) de las que dependen las aplicaciones. Los recursos de aplicación y red forman una unidad básica que gestiona RGM.

Grupos de recursos

Los recursos gestionados por RGM se colocan en grupos de recursos, de modo que se puedan gestionar como una unidad. Un grupo de recursos es un conjunto de recursos interdependientes o relacionados. Por ejemplo, un recurso derivado de un tipo de recurso `SUNW.LogicalHostname` es posible que se sitúe en el mismo grupo de recursos que el derivado de un tipo de recurso de base de datos de Oracle. El grupo de recursos migra como unidad si se inicia una recuperación de fallos o un cambio.

Tipos de servicios de datos

Los servicios de datos permiten a las aplicaciones estar altamente disponibles y los servicios escalables ayudan a evitar la interrupción significativa de las aplicaciones después de un único fallo en el clúster.

Cuando se configura un servicio de datos, se debe configurar como:

- Servicio de datos a prueba de fallos
- Servicio de datos escalable
- Servicio de datos paralelo

Servicios de datos a prueba de fallos

La recuperación de fallos es un proceso por el cual el clúster reubica automáticamente una aplicación de un nodo principal con fallos en un nodo designado secundario y redundante. Las aplicaciones a prueba de fallos tienen las características siguientes:

- Pueden ejecutarse en un solo nodo del clúster
- No están preparadas para el uso de clústers
- Dependen de la estructura del clúster para conseguir una alta disponibilidad

Si el supervisor de fallos detecta un error, intenta reiniciar la instancia del mismo nodo o la de otro nodo (recuperación de fallos), dependiendo de la forma en que se haya configurado el servicio de datos. Los servicios a prueba de fallos usan un grupo de recursos a prueba de fallos: un contenedor para los recursos de instancias de aplicaciones y recursos de la red (nombres de sistemas lógicos). Éstos son direcciones IP que pueden configurarse como activas en un nodo y, posteriormente, configurarse automáticamente como inactivas en el nodo original y activarse en otro nodo.

Es posible que los clientes hayan sufrido una breve interrupción del servicio y es posible que necesiten volver a conectarse después de terminar la recuperación del fallo. Sin embargo, los clientes no parecen estar al tanto del cambio en el servidor físico que proporciona el servicio.

Servicios de datos escalables

El servicio de datos escalables permite a las instancias de la aplicación ejecutarse en varios nodos de manera simultánea y usan dos grupos de recursos. El grupo de recursos escalable contiene los recursos de las aplicaciones y el grupo de recursos a prueba de fallos contiene los recursos de la red (direcciones compartidas), del que depende el servicio escalable. El grupo de recursos escalable puede estar en línea en varios nodos, de forma que se puedan ejecutar varias instancias del servicio simultáneamente. El grupo de recurso a prueba de fallos que aloja la dirección compartida está disponible en un solo nodo cada vez. Todos los nodos que alojan un servicio escalable usan la misma dirección compartida para alojar el servicio.

El clúster recibe las peticiones de servicio a través de una única interfaz de red (la interfaz global). Estas solicitudes se distribuyen a los nodos, basándose en uno de los algoritmos predefinidos especificados por la norma de equilibrado de cargas que el clúster puede usar para equilibrar la carga del servicio entre varios nodos.

Aplicaciones paralelas

Los sistemas Sun Cluster proporcionan un entorno que comparte la ejecución paralela de aplicaciones en todos los nodos del clúster, mediante las bases de datos paralelas. Admisión de Sun Cluster para Parallel Server/Real Application Clusters de Oracle es un conjunto de paquetes que, tras su instalación, permite la ejecución de Oracle Parallel Server/Real Application Clusters en nodos de Sun Cluster. Este servicio de datos también permite que las órdenes de Sun Cluster puedan gestionar Admisión de Sun Cluster para Parallel Server/Real Application Clusters de Oracle.

Una aplicación paralela se instrumentaliza de manera que se pueda ejecutar en un entorno del clúster y la puedan controlar simultáneamente dos o más nodos. En un entorno Oracle Parallel Server/Real Application Clusters, las diversas instancias de Oracle colaboran para proporcionar acceso a la misma base de datos compartida. Los clientes de Oracle pueden usar cualquiera de las instancias para acceder a la base de datos. De este modo, si una o mas instancias fallan, los clientes pueden conectarse con una instancia superviviente y continuar accediendo a la base de datos.

Arquitectura de Sun Cluster

La arquitectura de Sun Cluster permite el desarrollo, la gestión y la visualización de un grupo de sistemas como un único sistema grande.

Este capítulo se divide en los siguientes apartados:

- “Entorno de hardware de Sun Cluster” en la página 29
- “Entorno del software de Sun Cluster” en la página 30
- “Servicios de datos escalables” en la página 33
- “Almacenamiento en discos multisistema” en la página 35
- “Interconexión del clúster” en la página 35
- “Grupos de ruta múltiple de red IP” en la página 37

Entorno de hardware de Sun Cluster

Los componentes siguientes de hardware componen un clúster:

- Los nodos del clúster con discos locales (sin compartir) proporcionan la principal plataforma de computación del clúster.
- El almacenamiento multisistema proporciona discos compartidos entre los nodos.
- Los soportes extraíbles se configuran como dispositivos globales, como cintas y CD-ROM.
- La interconexión del clúster proporciona un canal para la comunicación intermodal.
- Las interfaces de red pública permiten que las interfaces de red que utilicen los sistemas de clientes tengan acceso a los servicios de datos del clúster.

La Figura 3–1 ilustra cómo trabajan conjuntamente los componentes del hardware.

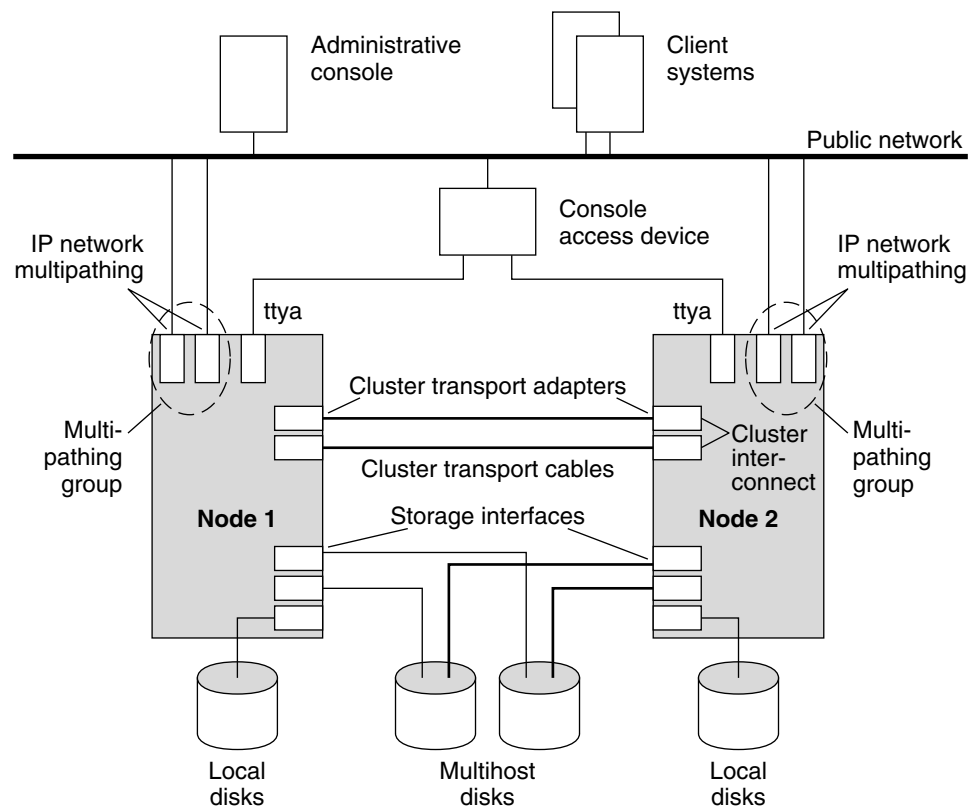


FIGURA 3-1 Componentes del hardware de Sun Cluster

Entorno del software de Sun Cluster

Si desea trabajar como miembro de un clúster, un nodo debe tener el siguiente software instalado:

- Software Solaris
- Software de Sun Cluster
- Aplicación de servicio de datos
- Gestión de volúmenes (Solaris™ Volume Manager o VERITAS Volume Manager)

Una excepción es una configuración que se suministre con una gestión de volúmenes. Es posible que no necesite un gestor de volúmenes de software.

La Figura 3-2 muestra con precisión cómo los componentes de software trabajan conjuntamente para crear el entorno de Sun Cluster.

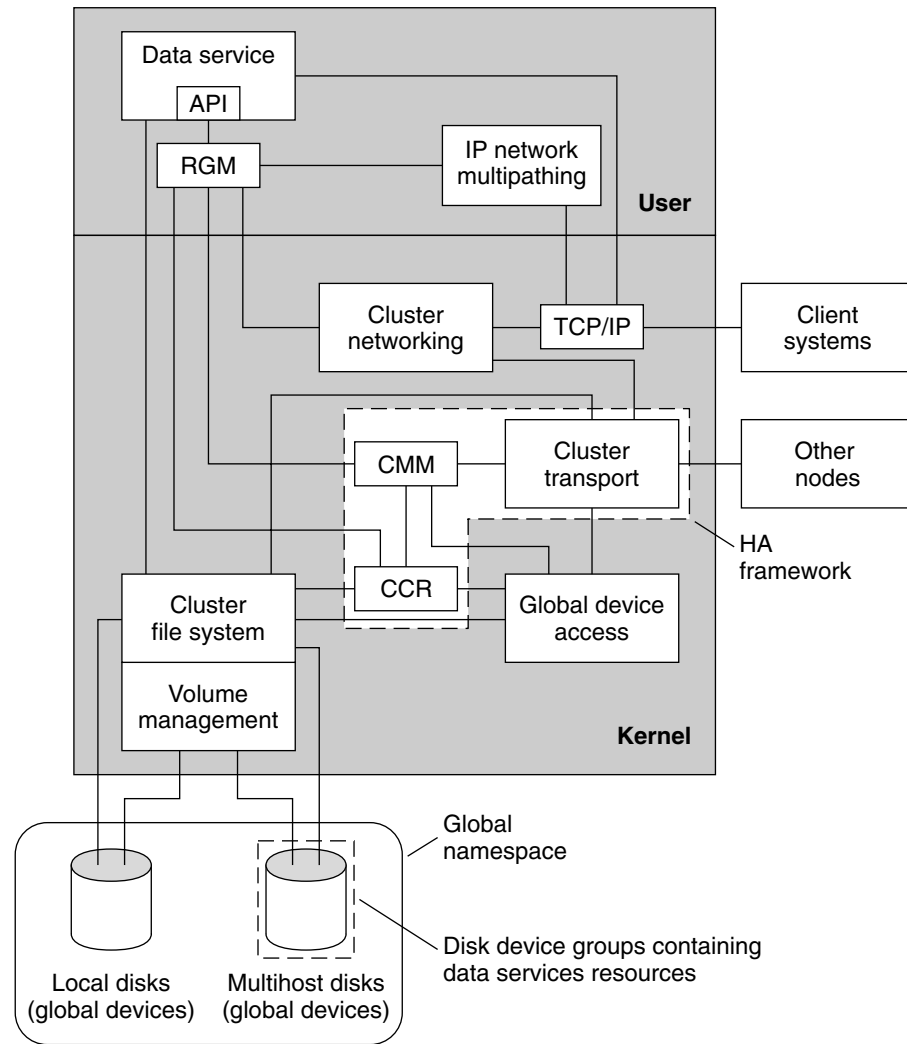


FIGURA 3-2 Arquitectura del software de Sun Cluster

Supervisor de pertenencia al clúster (CMM)

Para asegurarse de que los datos no sufran daños, todos los nodos deben alcanzar un acuerdo uniforme sobre la pertenencia al clúster. Cuando es necesario, CMM coordina una reconfiguración de los servicios del clúster en respuesta a un fallo.

CMM recibe información sobre conectividad con otros nodos desde la capa de transporte del clúster y usa la interconexión del clúster para intercambiar información de estado durante la reconfiguración.

Tras detectar un cambio en la pertenencia del clúster, CMM efectúa una configuración sincronizada de éste en la cual es posible que los recursos del clúster se redistribuyan, basándose en la nueva pertenencia del clúster.

CMM se ejecuta por completo en el núcleo.

Depósito de configuración del clúster (CCR)

CCR confía en CMM para garantizar que el clúster sólo se ejecute cuando se tenga el suficiente quórum. CCR es responsable de verificar la uniformidad de los datos entre el clúster, efectuando recuperaciones según sea necesario y facilitando actualizaciones a los datos.

Sistemas de archivos del clúster

Un sistema de archivos del clúster es un servidor proxy entre:

- El núcleo de un nodo y el sistema de archivos subyacente
- El gestor de volúmenes que se ejecute en un nodo con una conexión física con los discos

Los sistemas de archivos del clúster dependen de los dispositivos globales (discos, cintas, CD-ROM), accesibles desde cualquier nodo del clúster, a través del mismo nombre de archivo, (por ejemplo `/dev/global/`). Ese nodo no necesita una conexión física con el dispositivo de almacenamiento. Se puede utilizar un dispositivo global como dispositivo regular, esto es, se puede crear un sistema de archivos en un dispositivo global mediante `newfs` o `mkfs`.

El sistema de archivos del clúster ofrece las prestaciones siguientes:

- Las ubicaciones de los accesos de archivo son transparentes. Un proceso puede abrir un archivo que se encuentre en cualquier ubicación del sistema. Asimismo, los procesos de todos los nodos pueden usar el mismo nombre de ruta para ubicar un archivo.

Nota – Cuando el sistema de archivos del clúster lee archivos, no actualiza la hora de acceso en éstos.

- Se utilizan protocolos de coherencia para preservar la semántica de acceso a archivos UNIX aunque varios nodos estén accediendo al archivo al mismo tiempo.

- Para mover datos de archivos eficientemente se utiliza masivamente la antememoria y el movimiento de E/S en bloque sin copia.
- El sistema de archivos del clúster ofrece el bloqueo de archivos a través de las interfaces `fcntl(2)`. Las aplicaciones que se ejecuten en varios nodos del clúster pueden sincronizar el acceso a los datos mediante el bloqueo a los archivos de aviso en un archivo del sistema de archivos del clúster. Los bloqueos de archivo se recuperan inmediatamente desde los nodos que abandonan el clúster y las aplicaciones que fallan mientras se mantienen los bloqueos.
- El acceso continuo a los datos queda asegurado aunque se produzcan fallos. Las aplicaciones no se ven afectadas por fallos si sigue estando operativa una ruta de acceso a los discos. Esta garantía se mantiene para el acceso a discos de bajo nivel y todas las operaciones del sistema de archivos.
- Los sistemas de archivos del clúster son independientes del sistema de archivos subyacente y del software de gestión de volúmenes. Los sistemas de archivos del clúster convierten en global cualquier sistema de archivos admitido del disco.

Servicios de datos escalables

El objetivo principal de la conexión en red por clúster es ofrecer escalabilidad a los servicios de datos. Esto significa que a medida que aumente la carga ofrecida a un servicio, éste pueda mantener un tiempo de respuesta constante frente a este aumento de carga de trabajo según se vayan añadiendo nodos nuevos al clúster y se ejecuten instancias nuevas de servidores. Un buen ejemplo es un servicio web. Normalmente, un servicio de datos escalable se compone de varias instancias cada una de las cuales se ejecuta en distintos nodos del clúster. Cuando están juntas, estas instancias se comportan como un único servicio de un cliente remoto de ese servicio e implementa la funcionalidad del servicio. Un servicio web escalable con varios daemons `httpd` que se ejecuten en varios nodos puede hacer que cualquier daemon atienda las peticiones de un cliente. El que sirve la solicitud depende de una *norma de equilibrio de cargas*. La respuesta al cliente parece provenir del servicio, no del daemon concreto que atendió la petición, preservando así la apariencia de servicio individual.

La figura siguiente muestra la arquitectura de servicio escalable.

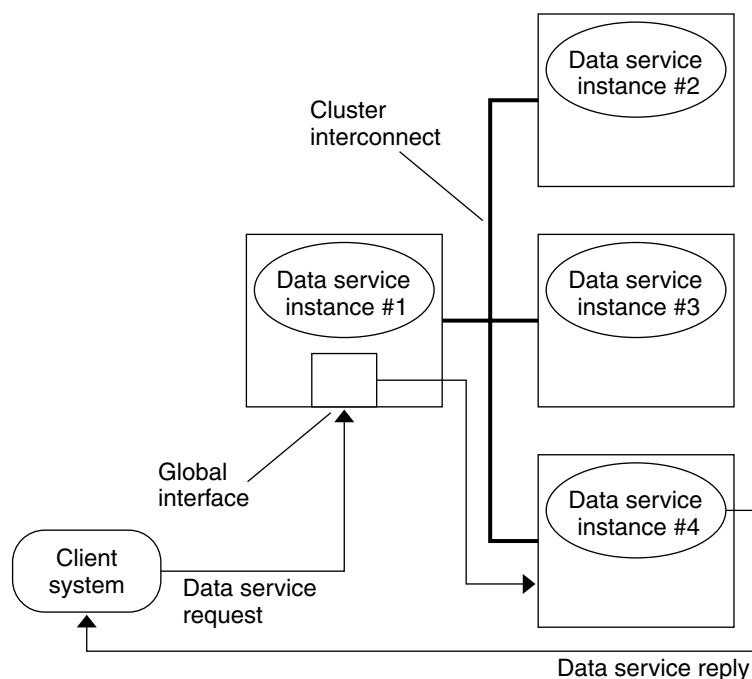


FIGURA 3-3 Arquitectura de servicio de datos escalable

Los nodos que no alojan la interfaz global (nodos delegados) tienen la dirección compartida alojada en sus interfaces de bucle. Los paquetes que se reciben en la interfaz global se distribuyen en otros nodos del clúster, basándose en las normas configurables de equilibrio de cargas. Las normas de equilibrio de cargas posibles se describen a continuación.

Normas de equilibrio de cargas

El equilibrio de cargas mejora el rendimiento del servicio escalable, tanto en tiempo de respuesta como en rendimiento.

Hay dos clases de servicios de datos escalables: *puros* y *adosados*. Un servicio puro es aquel donde una instancia puede responder a peticiones de clientes sin restricciones. Un servicio adosado tiene al clúster equilibrando la carga en las solicitudes al nodo. Estas peticiones no se redirigen a otras instancias.

Un servicio puro usa una política de equilibrio de cargas ponderada bajo la cual, predeterminadamente, las peticiones de los clientes se distribuyen de manera uniforme entre las instancias del servidor en el clúster. Por ejemplo, en un clúster de

tres nodos donde cada uno tenga el peso de 1, cada nodo atiende a un tercio de las solicitudes de cualquier cliente en nombre de ese servicio. Los pesos se pueden cambiar en cualquier momento mediante la interfaz de la orden `scrgadm(1M)` o mediante la interfaz de SunPlex Manager.

Un servicio adosado tiene dos tipos: *servicio adosado normal* y *servicio adosado comodín*. Ambos permiten que sesiones simultáneas de aplicación a través de varias conexiones TCP compartan el estado de la memoria (estado de la sesión de aplicación).

Los normales permiten a un cliente compartir el estado entre varias conexiones TCP simultáneas. Se considera que el cliente es “adosado” con respecto a la instancia del servidor que recibe en un único puerto. Al cliente se le garantiza que todas sus solicitudes vayan a la misma instancia del servidor, siempre que ésta permanezca activa y accesible y que la política de equilibrio de cargas no cambie mientras el servicio esté en línea.

Los servicios adosados comodín usan números de puerto asignados dinámicamente, pero siguen esperando que las peticiones de clientes vayan al mismo nodo. El cliente está “adosado con comodín” a los puertos con respecto a la misma dirección IP.

Almacenamiento en discos multisistema

El software Sun Cluster consigue de los discos una alta disponibilidad mediante el almacenamiento en discos multisistema, que se pueden conectar a más de un nodo a la vez. El software de gestión de volúmenes se puede utilizar para ordenar estos discos en un almacenamiento compartido controlado por un nodo del clúster. Los discos se configuran después para desplazarse a otro nodo si se produce un fallo. El uso de discos multisistema en sistemas Sun Cluster proporciona muchas ventajas, entre las que se encuentran:

- Acceso global a los sistemas de archivos
- Rutas de acceso múltiple a los datos y a los sistemas de archivos
- Tolerancia a los fallos de un único nodo

Interconexión del clúster

Todos los nodos deben estar conectados por la interconexión del clúster a través de, al menos, dos redes independientes físicamente o rutas de acceso, para evitar que exista un único punto de fallo. Puesto que son necesarias dos interconexiones para conseguir una mayor seguridad, se pueden usar hasta seis para dispersar el tráfico y evitar de

este modo los atascos, con lo que se mejora la seguridad y la escalabilidad. La interconexión del Sun Cluster utiliza Fast Ethernet, Gigabit-Ethernet, Sun Fire Link o Scalable Coherent Interface (SCI, IEEE 1596-1992), con lo que se consiguen comunicaciones privadas de alto rendimiento entre clústers.

En los entornos de clústers son esenciales las interconexiones de alta velocidad y baja latencia, así como los protocolos para comunicaciones intermodales. La interconexión SCI en los sistemas Sun Cluster ofrece un rendimiento mejorado en tarjetas estándar de interfaces de red (NIC). Sun Cluster utiliza la interfaz de memoria compartida remota (RSM™) en la comunicación intermodal en una red Sun Fire Link. RSM es una interfaz de mensajería de Sun altamente eficaz en operaciones de memoria remota.

La interconexión del clúster consta de los siguientes componentes de hardware:

- *Adaptadores*: las tarjetas de interfaz de red que residen en cada nodo de clústers. Un adaptador de red con varias interfaces podría convertirse en un único punto de fallo si fallara todo el adaptador.
- *Uniones*: los interruptores que se encuentran en el exterior de los nodos del clúster. Las uniones llevan a cabo funciones de transporte y conmutación para permitir conectar más de dos nodos simultáneamente. En un clúster de dos nodos no se necesitan las uniones puesto que los nodos se pueden conectar directamente entre sí mediante cables físicos redundantes que están conectados con adaptadores redundantes en cada nodo. Las configuraciones de más de dos nodos requieren uniones.
- *Cables*: las conexiones físicas que se sitúan entre dos adaptadores de red o un adaptador y una unión.

La Figura 3-4 muestra cómo se conectan los tres componentes.

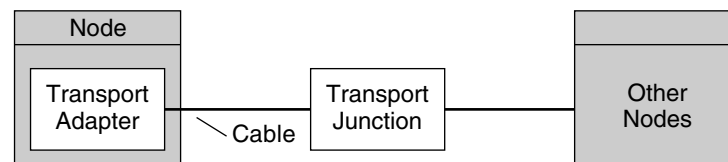


FIGURA 3-4 Interconexión del clúster

Grupos de ruta múltiple de red IP

Los adaptadores de red pública se organizan en grupos de ruta múltiple IP (grupos de ruta múltiple), cada uno de ellos con uno o más adaptadores de red pública. Cada adaptador de un grupo de ruta múltiple puede estar activo o se pueden configurar interfaces en espera que estén inactivos a menos que ocurra una recuperación de fallos.

Los grupos de ruta múltiple ofrecen la base para la construcción de nombres de sistema lógicos y recursos de dirección compartida. El mismo grupo de ruta múltiple de un nodo puede alojar cualquier número de nombres de sistema lógicos o recursos de dirección compartida. Si desea supervisar la conectividad de redes públicas de nodos del clúster, puede crear rutas múltiples.

Si desea más información sobre sistemas lógicos y recursos de direcciones compartidos, consulte *Sun Cluster Data Services Planning and Administration Guide for Solaris OS*.

Interfaces de red pública

Los clientes se conectan al clúster a través de interfaces de red pública. Todas las tarjetas adaptadoras de red pueden conectarse a una o más redes públicas, de acuerdo con las interfaces de hardware que la tarjeta tenga. Los nodos pueden configurarse para que incluyan múltiples tarjetas de interfaz de red pública a fin de que varias estén activas y se utilicen en caso de recuperación de fallos como respaldo entre ellas. Si uno de los adaptadores falla, el software de ruta múltiple de red del protocolo de Internet (IP) de Solaris, en Sun Cluster, recibe la instrucción de recuperarse de una interfaz defectuosa en otro adaptador del grupo.

Índice

A

- a prueba de fallos, servicios de datos, 27
- adaptadores, *Ver* adaptadores, red
- administración, herramientas, 14
- Admisión de Sun Cluster para Parallel Server/Real Application Clusters de Oracle, 28
- agentes, *Ver* servicios de datos
- aislamiento, 22
- almacenamiento
 - gestión, 12
 - matriz, 13
 - multisistema, 12, 35
- almacenamiento multisistema, 12
- alta disponibilidad, 9
- amnesia, 21
- API de gestión de recursos (RMAPI), 25
- API de la biblioteca de desarrollo del servicio de datos (DSDL API), 25
- aplicaciones
 - Ver también* servicios de datos
 - alta disponibilidad, 9
 - paralelas, 11, 28
 - supervisión, 13
 - tolerancia a fallos, 9
- aviso grave, 23

B

- bases de datos, 11
- bloqueo de archivo, 33

C

- clústers
 - comunicación, 18
 - configuración, 19, 32
 - interconexiones, 18, 35
 - miembros, 17, 31
 - nodos, 17
 - partición, 21
 - pertenencia, 18
 - red pública, 37
 - sistemas de archivos, 13, 32
- componentes
 - hardware, 29
 - software, 30
- configuración
 - base de datos paralela, 17
 - depósito, 19, 32
 - herramientas, 14
- conflicto de reserva, 23
- control del acceso, 15
- control del acceso basado en el rol (RBAC), 15
- controladores, *Ver* dispositivos, ID (DID)

D

- depósito, 19, 32
- Depósito de configuración del clúster (CCR), 32
- Depósito de configuración del clúster (CCR, Cluster Configuration Repository), 19
- dirección compartida, servicios de datos
 - escalables, 27

- discos
 - aislamiento de fallos, 22
 - dispositivos globales, 23
 - duplicación, 12, 13
 - gestión, 12
 - grupos de dispositivos, 25
 - locales, 23
 - multisistema, 12, 23, 25, 35
 - quórum, 21
- dispositivos
 - globales, 23
 - grupos, 25
 - ID (DID), 24
 - locales, 24
 - quórum, 21
- dispositivos globales
 - descripción, 23
 - grupos de dispositivos de discos, 25
 - montaje, 32
- dispositivos locales, 24

E

- entorno
 - hardware, 29
 - software, 30
- equilibrio de cargas
 - descripción, 33
 - normas, 34
- escalabilidad, *Ver* escalable
- escalables
 - grupos de recursos, 27
 - servicios, 11
 - servicios de datos, 27
 - arquitectura, 33
- espacio de nombres global, 24
- esquizofrenia, 21, 22

F

- fallo
 - aislamiento, 22
 - hardware y software, 13
- fallos, detección, 13

G

- gestión de disponibilidad, 10
- gestión de volúmenes, 12
- gestor de tráfico, 12
- Gestor del grupo de recursos (RGM)
 - funcionalidad, 25
 - grupos de recursos y, 26
- gestión de volúmenes, 35
- grupos de discos compartidos, 28

H

- hardware
 - alta disponibilidad, 9
 - entorno, 29
 - fallo, 13
 - interconexiones del clúster, 35
 - matriz redundante de discos independientes (RAID), 13
 - nodos del clúster, 17
 - Sun StorEdge Traffic Manager, 12
- herramientas, 14

I

- ID, dispositivo, 24
- integridad de los datos, 21
- interconexiones, *Ver* clúster, interconexiones
- interfaces, 20, 37
- interfaz de línea de órdenes (CLI), 15
- IPMP
 - Ver* Ruta múltiple de red IP

M

- matriz redundante de discos independientes (RAID), 13
- montaje, 32

N

- nodos, 17
- nombre de sistema lógico, servicios de datos a
 - prueba de fallos, 27

O

Oracle Parallel Server/Real Application Clusters, 12

P

paralelas

- aplicaciones, 11, 28
- bases de datos, 11, 17
- partición, clústers, 21
- pertenencia, 17, 18, 31
- Protocolo de Internet (IP), 28

Q

quórum, 21

R

- recuento de votos, quórum, 21
- recuperación, 9, 10
- recuperación de fallos
 - aportación del software Oracle Parallel Server/Real Application Clusters, 28
 - servicios, 11
 - transparente, 10
- recuperación rápida, 23
- recursos
 - compartidos, 13
 - definición, 26
 - grupos
 - a prueba de fallos, 27
 - descripción, 26
 - recuperación, 10
 - tipos, 26
- red
 - adaptadores, 11, 20, 37
 - equilibrio de cargas, 33, 34
 - interfaces, 11, 37
 - pública
 - descripción, 37
 - Ruta múltiple de red IP, 11, 20, 37
 - supervisión, 13
- red pública, *Ver* pública, red

redundancia

- hardware, 13
- sistemas de discos, 9
- ruta múltiple, 11, 13, 20, 37
- Ruta múltiple de red IP, 11, 20, 37

S

- scdpm, orden, 20
- SCSI, 22, 23
- servicios, *Ver* servicios de datos
- servicios de datos
 - a prueba de fallos, 27
 - definición, 25
 - escalables
 - adosados, 34
 - arquitectura, 33
 - puro, 34
 - recursos, 27
 - grupos de recursos, 26
 - paralelos, 28
 - recursos, 26
 - supervisión de fallos, 13
 - tipos, 26
 - tipos de recursos, 26
- sistemas de archivos
 - clústers, 13, 32
 - montaje, 32
- software
 - alta disponibilidad, 9
 - basado en el sistema, 13
 - componentes, 30
 - fallo, 13
 - matriz redundante de discos independientes (RAID), 13
- Solaris Volume Manager, 12
- Sun Management Center, 15
- Sun StorEdge Traffic Manager, 12
- SunPlex Manager, 14, 20
- supervisión
 - fallo, 13
 - interfaces de red, 20
 - rutas de discos, 20
- supervisión de rutas de discos (DPM), 20
- Supervisor de pertenencia al clúster (CMM), 18, 31

T

tolerancia a fallos, 9

V

VERITAS Volume Manager (VxVM), 12