Sun Cluster 3.1: Guía de conceptos

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

Referencia: 817–4259–10
Noviembre 2003, Revisión A

040112@7518

Contenido

Prefacio

Sun™ Sun Cluster 3.1: Guía de conceptos contiene información de conceptos y referencias sobre el sistema SunPlex™; incluye todos los componentes de hardware y software que componen la solución clúster de Sun.

Este documento está pensado para administradores de sistema con experiencia que conozcan el software Sun Cluster, no se debe usar como guía de preventa o de planificación. Antes de leerlo, debe conocer su sistema y disponer del equipo y el software adecuados.

Para entender los conceptos que se describen en este manual es necesario conocer el sistema operativo Solaris™ y tener expeciencia con el software de gestión de volúmenes que utiliza el sistema SunPlex.

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`	Nombres de los comandos, archivos y directorios; salida por pantalla del computador.	Edite el archivo `.login`. Utilice `ls -a` para mostrar una lista de todos los archivos. `nombre_sistema% 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 pantalla del computador	`nombre_sistema%` **`su`** `Password:`
AaBbCc123	Plantilla de línea de comandos: sustituir por un valor o nombre real	Para suprimir un archivo, escriba **`rm`** *nombre_archivo*.
AaBbCc123	Títulos de los manuales, palabras o términos nuevos o palabras destacables.	Véase el capítulo 6 de la *Guía del usuario* Se denominan opciones de *clase*. Para hacer esto debe ser el usuario *root*.

Indicadores de los shells en ejemplos de comandos

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	`#`

Documentación relacionada

Tema	Título	Número de referencia
Instalación	*Sun Cluster 3.1 10/03: Guía de instalación del software*	817–4253-10
Hardware	*Sun Cluster 3.x Hardware Administration Manual*	817-0168
	Colección Sun Cluster 3.x Hardware Administration en `http://docs.sun.com/db/coll/1024.1/`	
Servicios de datos	*Sun Cluster 3.1 Data Service Planning and Administration Guide*	817–3305
	Colección Sun Cluster 3.1 Data Services 10/03 en `http://docs.sun.com/db/coll/573.11`	
Desarrollo de API	*Sun Cluster 3.1 10/03: Guía del desarrollador de los servicios de datos*	817–4265-10
Administración	*Sun Cluster 3.1 10/03: Guía de administración del sistema*	817–4247-10
Mensajes de error	*Sun Cluster 3.1 10/03 Error Messages Guide*	817–0521
Páginas de comando man	*Sun Cluster 3.1 10/03 Reference Manual*	817–0522
Notas sobre la versión	*Sun Cluster 3.1 10/03: Notas sobre la versión*	817–4854–10
	Sun Cluster 3.x Release Notes Supplement	816-3381

Acceso a la documentación de Sun en línea

La sede web docs.sun.com[SM] 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 conocer una lista de documentos y cómo pedirlos, consulte "Adquirir documentación impresa" en http://docs.sun.com.

Obtención de ayuda

Si tiene problemas durante la instalación o utilización de SunPlex, 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, Sun Cluster 3.1)

Use los comandos siguientes para reunir información sobre todos los nodos del sistema para el proveedor de servicio.

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

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

CAPÍTULO 1

Introducción y visión general

El sistema SunPlex es una solución integrada de hardware y software Sun Cluster que se utiliza para crear servicios de alta disponibilidad y escalabilidad.

El manual Sun Cluster 3.1 10/03: Guía de conceptos proporciona la información conceptual que necesitan los usuarios principales de la documentación de SunPlex, entre los que se incluyen:

- Proveedores de servicio que instalan y reparan hardware clúster
- Administradores de sistemas que instalan, configuran y administran software de Sun Cluster
- Desarrolladores de aplicaciones que desarrollan servicios de recuperación de fallos y escalables para aplicaciones no incluidas actualmente en el producto Sun Cluster

Este manual complementa el resto de la documentación de SunPlex para proporcionar una visión completa del sistema SunPlex.

Este capítulo

- Proporciona una introducción y una vivión general del sistema SunPlex
- Describe los distintos puntos de vista de la audiencia de SunPlex
- Identifica conceptos clave que es necesario entender antes de trabajar con el sistema SunPlex
- Correlaciona los conceptos clave con la documentación de SunPlex que incluye procedimientos e información relacionada
- Correlaciona tareas relacionadas del clúster con la documentación que contiene procediminentos usados para realizar esas tareas

Introducción al sistema SunPlex

El sistema SunPlex amplía el sistema operativo Solaris hasta hacerlo un sistema operativo de clúster. Un clúster, o plex, es una colección de nodos informáticos acoplados indirectamente que proporcionan una vista de cliente único de servicios de red o de aplicaciones, incluidos bases de datos, servicios de web y servicios de archivos.

Cada nodo del clúster es un servidor autónomo que ejecuta sus propios procesos. Éstos se comunican entre sí para formar lo que parece (a un cliente de red) como un sólo sistema que proporciona de forma cooperativa aplicaciones, recursos de sistema y datos a usuarios.

Un clúster ofrece, respecto a los sistemas tradicionales de servidor único, varias ventajas que incluyen soporte para servicios de protección contra fallos y escalabilidad, capacidad para crecimiento modular y un precio básico económico en comparación a los sistemas de hardware tolerantes a fallos.

Los objetivos del sistema SunPlex son:

- Reducir o eliminar el tiempo de inactividad del sistema debido a los fallos de software o hardware
- Asegurar la disponibilidad de los datos y aplicaciones a los usuarios finales, cualquiera que sea el tipo de fallo que normalmente dejaría inactivo un sistema de servidor único
- Aumentar el rendimiento de las aplicaciones permitiendo escalar procesadores adicionales con solo agregar más nodos al clúster
- Proporcionar una mejor disponiblidad del sistema al permitirle realizar el mantenimiento sin tener que apagar todo el clúster

Alta disponibilidad frente a tolerancia a fallos

El sistema SunPlex está diseñado como sistema de *alta disponibilidad* (AD), es decir, capaz de proporcionar acceso casi indefinido a datos y aplicaciones.

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. Además, los sistemas tolerantes a fallos normalmente no tienen en cuenta los fallos del software.

El sistema SunPlex consigue gran disponibilidad a través de una combinación de hardware y software. El clúster redundante interconecta y almacena mientras las redes públicas protegen contra los puntos de fallo individuales. El software del clúster supervisa continuamente el buen funcionamiento de los nodos miembros y evita que

los que no funcionen participen en el clúster para protegerlo frente a la corrupción de datos. Además, el clúster supervisa los servicios y los recursos del sistema dependientes y sustituye o reinicia los servicios en caso de anomalías.

Consulte «FAQ sobre alta disponibilidad» en la página 87 para consultar preguntas y respuestas referentes a la alta disponibilidad.

Recuperación de fallos y escalabilidad en el sistema SunPlex

El sistema SunPlex permite implementar servicios de *recuperación de fallos* o de *escalabilidad*. Por lo general, un servicio de recuperación de fallos sólo ofrece una alta disponibilidad (redundancia), mientras que un servicio de escalabilidad ofrece alta disponiblidad junto con un mayor rendimiento. Un clúster individual puede admitir ambos tipos de servicios.

Servicios de recuperación de fallos

La recuperación de fallos es el proceso por el que el clúster reubica automáticamente un servicio de un nodo primario fallido en un nodo secundario designado. Con la recuperación de fallos, el software Sun Cluster porporciona una alta disponibilidad.

Cuando actúa la recuperación de fallos, los clientes podrían experimentar una breve interrupción del servicio y es posible que tengan que reconectar cuando la recuperación haya finalizado. Sin embargo, para los clientes parece que sólo haya un único servidor que proporcione el servicio.

Servicios de escalabilidad

Mientras que la recuperación de fallos está relacionada con la redundancia, la escalabilidad ofrece un tiempo de respuesta o rendimiento constantes independientemente de la carga. Un servicio escalable aprovecha los distintos nodos de un clúster para ejecutar aplicaciones concurrentemente, ofreciendo así un mayor rendimiento. En una configuración escalable, todos los nodos del clúster pueden proporcionar datos y procesar peticiones de los clientes.

Consulte «Servicios de datos» en la página 57 para obtener información específica sobre los servicios de recuperación de fallos y de escalabilidad.

Tres puntos de vista sobre el sistema SunPlex

Este apartado describe tres puntos de vista distintos sobre el sistema SunPlex, los conceptos clave y la documentación relevante para cada punto de vista. Estos puntos de vista provienen de:

- Personal de instalación y reparación de hardware
- Administradores del sistema
- Programadores de aplicaciones

Punto de vista del personal de instalación y reparación de hardware

Para los profesionales de la reparación del hardware, el sistema SunPlex es como una colección de hardware estándar que incluye servidores, redes y almacenamiento. Estos componentes están todos interconectados de manera que todos tengan un recambio y no exista un punto débil para los fallos.

Conceptos clave del hardware

El personal de reparación de hardware necesita entender los conceptos de clúster siguientes.

- Configuraciones de hardware de clúster y cableado
- Instalación y reparación (agregar, eliminar, sustituir):
  - Componentes de la interfaz de red (adaptadores, uniones, cables)
  - Tarjetas interfaz de disco
  - Matrices de disco
  - Unidades de disco
  - La consola de administración y el dispositivo de acceso a la consola
- Configuración de la consola de administración y del dispositivo de acceso a la consola

Referencias conceptuales de hardware resomendadas

Los apartados siguientes contienen material relacionado con los conceptos clave anteriores:

- «Nodos del clúster» en la página 20

Información destacable sobre SunPlex

El documento de SunPlex siguiente incluye procedimientos e información asociada con conceptos de servicios de hardware:

- *Sun Cluster 3.1 Hardware Collection*

Punto de vista del administrador de sistemas

Para el administrador, el sistema SunPlex se asemeja a un conjunto de servidores (nodos) interconectados y que comparten dispositivos de almacenamiento. El adiministrador del sistema ve:

- Software para clústers especializado que se integra con el de Solaris para supervisar la conectividad entre los nodos del clúster
- Software especializado que supervisa el buen funcionamiento de los programas de aplicación del usuario que se ejecutan en sus nodos del clúster
- Software de gestión de volúmenes que configura y administra discos
- Software de clúster especializado que permite a todos los nodos acceder a todos los dispositivos de almacenamiento, incluso a aquellos no conectados directamente a discos
- Software de clúster especializado que permite a los archivos aparecer en todos los nodos como si estuvieran alojados localmente en él

Conceptos clave de la administración de sistemas

Es importante que los administradores de sistemas comprendan los conceptos y procesos siguientes:

- La interacción entre los componentes de hardware y software
- El flujo general sobre cómo instalar y configurar el clúster, que incluye:
  - Instalación del sistema operativo Solaris
  - Instalación y configuración del software Sun Cluster

- Instalación y configuración de un gestor de volúmenes
- Instalación y configuración del software de aplicación para que esté preparado para el clúster
- Instalación y configuración del software del servicio de datos de Sun Cluster

- Procedimientos administrativos del clúster para agregar, eliminar, sustituir y reparar componentes de hardware y software del clúster
- Modificaciones de la configuración para mejorar el rendimiento

Referencias conceptuales recomendadas para administradores de sistemas

Los apartados siguientes contienen material relacionado con los conceptos clave anteriores:

- «Interfaces administrativas» en la página 34
- «Estructura de alta disponibilidad (HA)» en la página 36
- «Dispositivos globales» en la página 39
- «Grupos de dispositivos de disco» en la página 40
- «Espacio de nombres global» en la página 44
- «Sistemas de archivos del clúster» en la página 45
- «Quórum y dispositivos del quórum» en la página 52
- «Gestores de volúmenes» en la página 57
- «Servicios de datos» en la página 57
- «Recursos: definición, grupos y tipos» en la página 69
- «Adaptadores de red pública y Ruta múltiple de red IP» en la página 81
- Capítulo 4

Documentación de SunPlex importante para el administrador de sistemas

Los documentos de SunPlex siguientes incluyen procedimientos e información asociada a los conceptos de la administración de sistemas:

- *Sun Cluster 3.1: Guía de instalación del software*
- *Sun Cluster 3.1: Guía de administración del sistema*
- *Sun Cluster 3.1 Error Messages Guide*
- *Sun Cluster 3.1: Notas sobre la versión*
- *Sun Cluster 3.1 Release Notes Supplement*

Punto de vista del programador de aplicaciones

El sistema SunPlex proporciona *servicios de datos* para aplicaciones como Oracle, NFS, DNS, Sun™ ONE Web Server, Apache Web Server y Sun™ ONE Directory Server. Los servicios de datos se crean configurando aplicaciones estándar de forma que se ejecuten bajo el control del software Sun Cluster. Éste proporciona archivos de

configuración y métodos de gestión que inician, paran y supervisan las aplicaciones. Si necesita crear un nuevo servicio a prueba de fallos o escalable, puede usar la Interfaz de programación de aplicaciones (API) de SunPlex y la API de tecnologías de habilitación de servicio de datos (DSET API) para desarrollar los archivos de configuración y métodos de gestión necesarios que permitan a esa aplicación ejecutarse como servicio de datos en el clúster.

Conceptos clave del programador de aplicaciones

Es importante que los programadores de aplicaciones entiendan lo siguiente:

- Las caracteristicas de la aplicación, para determinar si ésta puede ejecutarse como servicio de recuperación de fallos o de escalabilidad de datos.
- La API de Sun Cluster, API DSET y el servicio de datos “genérico”. Los programadores deben determinar qué herramientas son más adecuadas para usarlas al escribir programas o secuencias que configuren su aplicación para el entorno clúster.

Referencias conceptuales recomendadas para los programadores de aplicaciones

Los apartados siguientes contienen material relacionado con los conceptos clave anteriores:

- «Servicios de datos» en la página 57
- «Recursos: definición, grupos y tipos» en la página 69
- Capítulo 4

Documentación sobre SunPlex importante para el programador de aplicaciones

Los documentos de SunPlex siguientes incluyen procedimientos e información asociada a los conceptos de la programación de aplicaciones:

- *Sun Cluster 3.1: Guía del desarrollador de los servicios de datos*
- *Sun Cluster 3.1 Data Services Installation and Configuration Guide*

Tareas del sistema de SunPlex

Todas las tareas del sistema SunPlex requieren algunos conocimientos generales previos. La tabla siguiente incluye una visión general de las tareas y la documentación que describe sus pasos. El apartado de conceptos de este manual decribe la correlación entre los conceptos y estas tareas.

TABLA 1–1 Mapa de tareas: correspondencia entre tareas de usuario y documentación

Para realizar esta tarea...	Usar esta documentación...
Instalar el hardware del clúster	*Sun Cluster 3.1 Hardware Collection*
Instalar el software de Solaris en el clúster	*Sun Cluster 3.1: Guía de instalación del software*
Instalar el software Sun™ Management Center	*Sun Cluster 3.1: Guía de instalación del software*
Instalar y configurar el software Sun Cluster	*Sun Cluster 3.1: Guía de instalación del software*
Instalar y configurar el software de gestión de volúmenes	*Sun Cluster 3.1: Guía de instalación del software* La documentación del gestor de volúmenes
Instalar y configurar los servicios de datos para Sun Cluster	*Sun Cluster 3.1 Data Services Installation and Configuration Guide*
Reparar el hardware del clúster	*Sun Cluster 3.1 Hardware Collection*
Administrar el software de Sun Cluster	*Sun Cluster 3.1: Guía de administración del sistema*
Administrar el software de la gestión de volúmenes	*Sun Cluster 3.1: Guía de administración del sistema* y la documentación de la gestión de volúmenes propia
Administrar el software de las aplicaciones	La documentación de su aplicación
Identificación de los problemas y las acciones de usuario sugeridas	*Sun Cluster 3.1 Error Messages Guide*
Crear un servicio de datos nuevo	*Sun Cluster 3.1: Guía del desarrollador de los servicios de datos*

CAPÍTULO 2

Conceptos clave de los proveedores de servicio de hardware

Este capítulo describe los conceptos clave relacionados con los componentes de hardware de una configuración de sistema de SunPlex. Se tratan los siguientes temas:

Los componentes de hardware del sistema SunPlex

Esta información está dirigida principalmente a los proveedores de servicio de hardware a quienes ayuda a entender la relación entre los componentes de hardware para mejor instalar, configurar o reparar el hardware del clúster. Asimismo, a los administradores del sistema del clúster esta información les puede resultar útil para instalar, configurar y administrar el software del clúster.

Un clúster consta de varios componentes de hardware, entre los que se incluyen:

- Nodos de clúster con discos locales (no compartidos)
- Almacenamiento multisistema (discos compartidos entre nodos)
- Soporte extraíble (cintas y CD-ROM)
- Interconexión del clúster

- Interfaces de red públicas
- Sistemas cliente
- Consola de administración
- Dispositivos de acceso a la consola

El sistema SunPlex permite combinar estos componentes en diversas configuraciones, que se describen en «Ejemplos de topología de Sun Cluster » en la página 28.

Esta ilustración muestra un ejemplo de configuración del clúster.

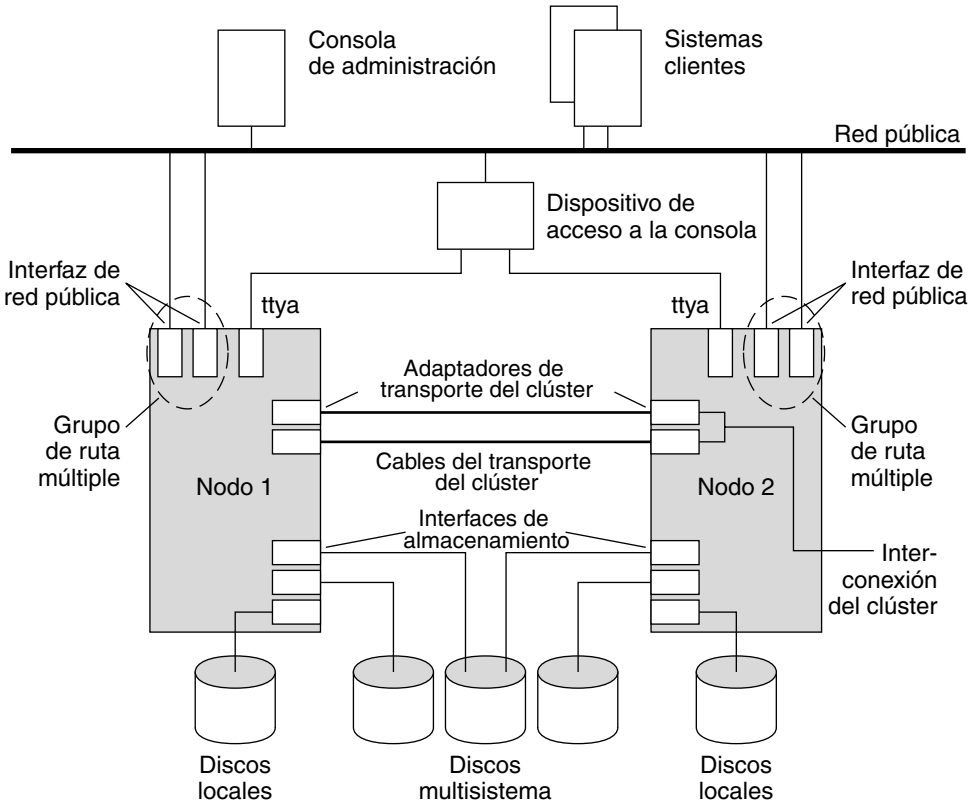


FIGURA 2–1 Configuración de clúster con dos nodos

Nodos del clúster

Un nodo del clúster es una máquina que ejecuta el sistema operativo Solaris y el sofware de Sun Cluster y es un componente actual del clúster (un *miembro del clúster*) o un miembro potencial. El software de Sun Cluster permite tener de dos a ocho nodos por clúster. Consulte en «Ejemplos de topología de Sun Cluster » en la página 28 las configuraciones de nodos admitidas.

Los nodos del clúster normalmente están conectados a uno o más discos multisistema, aquéllos que no lo están usan el sistema de archivos del clúster para acceder a los discos multisistema. Por ejemplo, una configuración de servicio escalable permite que los nodos atiendan peticiones sin estar directamente conectados a los discos multisistema.

Además, los nodos en configuraciones de bases de datos paralelas comparten acceso simultáneo a todos los discos. Consulte «Discos multisistema» en la página 23 y el Capítulo 3 para obtener más información sobre las configuraciones de bases de datos paralelas.

Todos los nodos del clúster están agrupados bajo un nombre común (el nombre del clúster) que se utiliza para acceder a éste y gestionarlo.

Los adaptadores de redes públicas conectan los nodos a éstas para ofrecer acceso de cliente al clúster.

Los miembros del clúster se comunican entre sí a través de una o más redes físicamente independientes que reciben el nombre de *interconexión del clúster*.

Todos los nodos del clúster reciben información de cuándo un nodo se une o deja el clúster, conocen también los recursos que se están ejecutando localmente así como los que se están ejecutando 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 recuperación de fallos, todos los nodos deben tener suficiente capacidad sobrante para que los que sean de respaldo o secundarios puedan aprovechar la carga de trabajo.

Cada nodo ejecuta su propio sistema de archivos raíz individual (/).

Componentes del software para los miembros del hardware del clúster

Para poder actuar como miembro del clúster, es necesario tener instalado el software siguiente:

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

  Una excepción es la configuración que usa una matriz redundante de hardware de discos independientes (RAID). Esta configuración puede no requerir un software de gestión de volúmenes del tipo Gestor de volúmenes de Solaris o VERITAS Volume Manager.

Consulte *Sun Cluster 3.1: Guía de instalación del software* para obtener información sobre cómo instalar el software del sistema operativo Solaris, Sun Cluster y el gestor de volúmenes.

Consulte *Sun Cluster 3.1 Data Services Installation and Configuration Guide* para obtener información sobre cómo instalar y configurar los servicios de datos.

Consulte el Capítulo 3 para obtener información básica sobre los componentes del software mencionados anteriormente.

La figura siguiente incluye una vista detallada de los componentes del software que funcionan juntos para crear el entorno Sun Cluster.

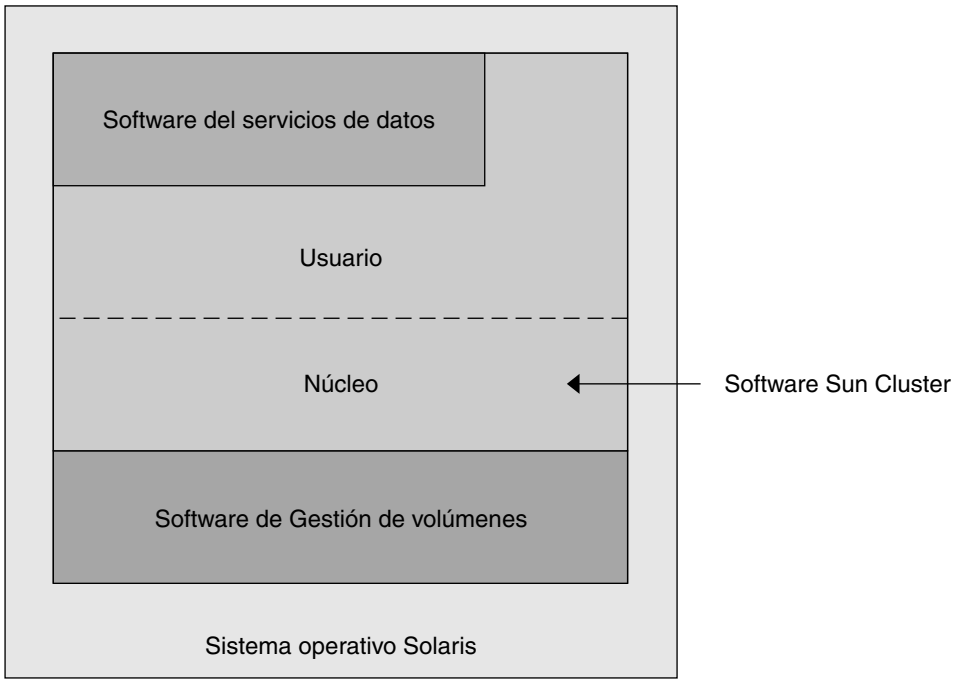


FIGURA 2–2 Relaciones entre los componentes del software de Sun Cluster

Consulte el Capítulo 4 para obtener preguntas y respuestas sobre los miembros de clúster.

Discos multisistema

Los discos que pueden conectarse a más de un nodo simultáneamente se denominan multisistema. En el entorno Sun Cluster, el almacenamiento multisistema hace que los discos estén muy disponibles. Sun Cluster requiere que haya almacenamiento multisistema en dos clústers del nodo para establecer el quórum. Los clústers de más de tres nodos no requieren almacenamiento multisistema.

Los discos multisistema poseen las características siguientes.

- Pueden resistir los fallos de nodos individuales.
- Almacenan datos de las aplicaciones y también pueden almacenar los archivos binarios y de configuración de éstas.
- Protegen contra los fallos del nodo. Si las peticiones de los clientes están accediendo a datos a través de un nodo y éste falla, las peticiones se desvían para usar otro nodo que tenga una conexión directa con los mismos discos.
- A ellos se accede tanto globalmente a través de un nodo primario que "coordina" los discos como por acceso directo simultáneo a través de las rutas de acceso locales. Actualmente la única aplicación que usa acceso directo simultáneo es Oracle Parallel Server.

Un gestor de volúmenes proporciona configuraciones especulares o RAID-5 para la redundancia de los discos multisistema. Actualmente, Sun Cluster admite el Gestor de volúmenes de Solaris™ y VERITAS Volume Manager como gestores de volúmenes y el controlador de hardware RAID-RD5 en varias plataformas de hardware RAID.

Al combinar los discos multisistema con los discos especulares y la separación de datos se consigue proteger contra el fallo de los nodos y el de los discos individuales.

Consulte el Capítulo 4 para obtener preguntas y respuestas sobre el almacenamiento multisistema.

Iniciador múltiple SCSI

Este apartado se aplica sólo a dispositivos de almacenamiento SCSI y no a los de fibra óptica que usan los discos multisistema.

En un servidor autónomo, el nodo servidor controla las actividades del bus SCSI porque el circuito adaptador del sistema SCSI se conecta a ese servidor en un bus SCSI determinado. Al circuito adaptador del sistema SCSI se le conoce como *iniciador SCSI*. Este circuito inicia todas las actividades de este bus SCSI. La dirección SCSI predeterminada de los adaptadores en el sistema Sun es 7.

Las configuraciones del clúster comparten el almacenamiento entre varios nodos del servidor, usando los discos multisistema. Cuando el almacenamiento del clúster se compone de dispositivos SCSI de terminación simple o diferencial a la configuración se la denomina iniciador múltiple SCSI. Tal como implica esta terminología, existe más de un iniciador SCSI en el bus.

La especificación SCSI requiere que cada dispositivo de un bus SCSI tenga una dirección exclusiva. (El sistema adaptador también es un dispositivo del bus.) La configuración de hardware predeterminada en un entorno de iniciador múltiple provoca un conflicto debido a que de forma predeterminada todos los adaptadores de sistema SCSI tienen el valor 7.

Para resolver este conflicto, se debe dejar uno de los adaptadores de cada bus SCSI con la dirección SCSI 7 y se debe configurar los otros adaptadores como direcciones SCSI no utlizadas. La planificación correcta dictamina que estas direcciones SCSI "no utilizadas" incluyan tanto direcciones usadas actualmente como no utilizadas. Un ejemplo de direcciones no utilizadas en el futuro es la incorporación del almacenamiento instalando unidades nuevas en ranuras de unidad vacías. En la mayoría de configuraciones, la dirección SCSI disponible para un segundo adaptador de sistema es 6.

Se pueden cambiar las direcciones SCSI seleccionadas de esos adaptadores configurando la propiedad `scsi-initiator-id` de la Open Boot PROM (OBP) tanto globalmente para un nodo como individualmente para cada adaptador. En el capítulo para cada alojamiento de disco de *Sun Cluster 3.1 Hardware Collection* hay instrucciones para configurar un `scsi-initiator-id` exclusivo para cada adaptador SCSI.

Discos locales

Los discos locales son aquellos que sólo estan conectados a un nodo individual. Por tanto, no están protegidos contra ningún fallo del nodo (no son de alta disponibilidad). A pesar de ello, todos los discos, incluso los locales se incluyen en el espacio de nombres global y se configuran como *dispositivos globales*. Por lo tanto, los discos en sí son visibles desde todos los nodos del clúster.

Puede poner los sistemas de archivos de discos locales a disposición de otros nodos situándolos bajo un punto de montaje global. Si el nodo que tiene montado actualmente uno de estos sistemas de archivos globales falla, el resto de nodos pierde acceso a ese sistema de archivos. Usar un gestor de volúmenes permite duplicar estos discos de forma que un fallo no pueda provocar que los sistemas de archivos dejen de estar accesibles, aunque los gestores de volúmenes no protejan contra los fallos de los nodos.

Consulte el apartado «Dispositivos globales» en la página 39 para obtener más información sobre dispositivos globales.

Soportes extraíbles

El clúster admite soportes extraíbles, como unidades de cinta y de CD-ROM, que se instalan, configuran y reparan de la misma forma que en un entorno que no está configurado en clúster. Estos dispositivos están configurados como globales en Sun Cluster, de manera que son accesibles desde cualquier nodo del clúster. Consulte *Sun Cluster 3.1 Hardware Collection* para obtener información sobre cómo instalar y configurar soportes extraíbles.

Consulte el apartado «Dispositivos globales» en la página 39 para obtener más información sobre dispositivos globales.

Interconexión del clúster

La *interconexión del clúster* es la configuración física de dispositivos que se utiliza para transferir comunicaciones privadas del clúster y de servicios de datos entre los nodos del clúster. Debido a que la interconexión se utiliza ampliamente para comunicaciones privadas del clúster, puede limitar el rendimiento.

La interconexión del clúster sólo puede conectar los distintos nodos del clúster. El modelo de seguridad Sun Cluster asume que sólo los nodos del clúster tienen acceso físico a la 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. Entre dos nodos cualesquiera puede tener varias redes independientes físicamente (de dos a seis). La interconexión del clúster se compone de tres componentes del hardware: adaptadores, uniones y cables.

La lista siguiente describe cada uno de estos componentes del hardware.

- Adaptadores: las tarjetas de interfaz de red que residen en cada nodo del clúster. Sus nombres se construyen a partir de un nombre de dispositivo seguido inmediatamente por un número de unidad física, por ejemplo: qfe2. Algunos adaptadores sólo tienen una conexión de red física, aunque otros como la tarjeta qfe tienen varias conexiones físicas. Algunos también contienen interfaces de red y de almacenamiento.

  Un adaptador de red con varias interfaces podría convertirse en un único punto de fallo si falla todo el adaptador. Para una máxima disponibilidad, se debe planificar el clúster de manera que la única ruta entre dos nodos no dependa de un único adaptador de red individual.
- Uniones: los conmutadores que residen fuera de los nodos del clúster. Llevan a cabo funciones de transporte y conmutación para permitir conectar más de dos nodos simultáneamente. En un clúster de dos nodos no hacen falta uniones porque los nodos pueden conectarse directamente entre sí con cables físicos redundantes conectados a adaptadores redundantes de cada nodo. Las configuraciones superiores a dos nodos en general requieren uniones.

- Cables: las conexiones físicas que van entre dos adaptadores de red o entre un adaptador y una unión.

Consulte el Capítulo 4 para obtener preguntas y respuestas sobre la interconexión del clúster.

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 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, se llama al software Ruta múltiple de red IP para la recuperación de la intefaz defectuosa pasando a otro adaptador del grupo.

No hay consideraciones de hardware especiales relacionadas con la agrupación en clúster para las interfaces de red públicas.

Consulte el Capítulo 4 para obtener preguntas y respuestas sobre las redes públicas.

Sistemas cliente

Los sistemas cliente son las estaciones de trabajo u otros servidores que acceden al clúster a través de la red pública. Los programas del lado del cliente usan datos u otros servicios proporcionados por aplicaciones del lado del servidor que se ejecutan en el clúster.

Los sistemas cliente no son de alta disponibilidad. Los datos y aplicaciones del clúster son de alta disponiblidad.

Consulte el Capítulo 4 para obtener preguntas y respuestas sobre los slstemas cliente.

Dispositivos de acceso a la consola

Es necesario que disponga de acceso a la consola para todos los nodos del clúster. Para ello, use el concentrador de terminal que ha adquirido con el hardware del clúster, el procesador de servicio del sistema (SSP) de los sevidores Sun Enterprise E10000™, el controlador del sistema en servidores Sun Fire™ u otro dispositivo que pueda acceder a `ttya` en todos los nodos.

Sun sólo dispone de un concentrador de terminal admitido y su uso es opcional. El concentrador de terminal permite el acceso a `/dev/console` en todos los nodos a través de una red TCP/IP. El resultado es el acceso a nivel de consola para todos los nodos desde una estación de trabajo remota desde cualquier lugar de la red.

El procesador de servicio del sistema (SSP) proporciona acceso a la consola para el servidor Sun Enterprise E10000. SSP es una máquina de una red Ethernet que está configurada para admitir el servidor Sun Enterprise E10000. SSP es la consola de administración para el servidor Sun Enterprise E10000. Gracias a la función Sun Enterprise E10000 Network Console, cualquier estación de trabajo de la red puede abrir una sesión de consola de sistema.

Otros métodos de acceso a la consola son concentradores de otras terminales, acceso de puerto serie por `tip(1)` desde otro nodo y terminales no inteligentes. Puede usar teclados y monitores de Sun™ u otros dispositivos de puerto serie si el proveedor de servicio de hardware los admite.

Consola de administración

Para administrar el clúster activo se puede emplear una estación de trabajo UltraSPARC™ exclusiva, conocida como *consola de administración* que normalmente se instala y ejecuta software de utilidad de administración como el panel de control del clúster (CCP) y el módulo de Sun Cluster para el producto Sun Management Center™. Utilizando `cconsole` en CCP se le permitirá conectar a más de un nodo simultáneamiente. Para obtener mas información sobre el uso de CCP, consulte *Sun Cluster 3.1: Guía de administración del sistema*.

La consola de administración normalmente no es un nodo del clúster. Se utiliza para acceder a los nodos del clúster de forma remota, ya sea a través de la red pública ya sea, opcionalmente, mediante un concentrador de terminales ubicado en la red. Si el clúster se compone de la plataforma Sun Enterprise E10000, es necesario que se pueda iniciar la sesión desde la consola de administración con el procesador de servicio del sistema (SSP) y que se pueda conectar mediante el comando `netcon(1M)`.

Normalmente los nodos se configuran sin monitor. A continuación se accede a la consola del nodo empleando una sesión `telnet` desde la consola de administración, que está conectada a un concentrador de terminal y desde éste al puerto serie del nodo. (En el caso del servidor Sun Enterprise E10000, se conecta desde el procesador de servicio del sistema.) Consulte «Dispositivos de acceso a la consola» en la página 26 para obtener más información.

Sun Cluster no requiere una consola de administración exclusiva, aunque si se usa una se obtendrán las siguientes ventajas:

- Permite la gestión centralizada del clúster ya que agrupa herramientas de consola y gestión en la misma máquina
- Ofrece al proveedor de servicio de hardware una resolución de problemas potencialmente mas rápida

Consulte el Capítulo 4 para obtener preguntas y respuestas sobre la consola de administración.

Ejemplos de topología de Sun Cluster

Una topología es el esquema de conexión que une los nodos del clúster con las plataformas de almacenamiento que se usan en éste. Sun Cluster admite cualquier topología que cumpla las siguientes directrices.

- Sun Cluster admite un máximo de ocho nodos por clúster, sean cuales sean las configuraciones de almacenamiento implantadas.
- Un dispositivo de almacenamiento compartido puede conectarse a tantos nodos como se admitan.
- Los dispositivos de almacenamiento compartido no necesitan conectarse con todos los nodos del clúster. Sin embargo, estos dispostivos de almacenamiento deben conectarse a un mínimo de dos nodos.

Sun Cluster no obliga a configurar el clúster mediante topologías específicas. Las topologias que se describen a continuación se incluyen para proporcionar el vocabulario que permita discutir un esquema de conexión del clúster. Estas topologías son esquemas de conexión típicos.

- Pares en clúster
- Par+N
- N+1 (estrella)
- N*N (escalable)

Los apartados siguientes incluyen diagramas de ejemplo para cada topología.

Topología de pares en clúster

Una topología de pares en clúster son dos o más pares de nodos que funcionan bajo un único marco administrativo de clúster. En esta configuración sólo se produce recuperación de fallos entre pares. Sin embargo, todos los nodos están conectados por la interconexión del clúster y funcionan bajo el control del software Sun Cluster. Esta topología se podría usar para ejecutar una aplicación de base de datos paralela en un par y una aplicación de recuperación de fallos o de escalabilidad en otro par.

Con el sistema de archivos del clúster, también podría tener una configuración de dos pares en que más de dos ejecuten un servicio escalable o base de datos paralela aunque ningún nodo esté conectado directamente a los discos que almacenan los datos de la aplicación.

La figura siguiente ilustra una configuración de par en clúster.

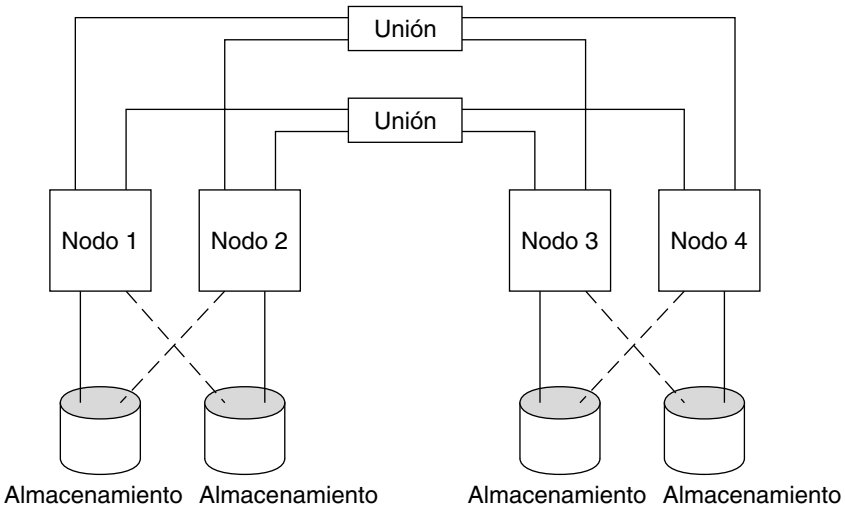


FIGURA 2–3 Topología de pares en clúster

Topología par+n

La topología par+n incluye un par de nodos conectados directamente al almacenamiento compartido y un conjunto adicional de nodos que usa la interconexión del clúster para acceder al almacenamiento compartido; éstos no tienen conexión directa.

La figura siguiente muestra una topología par+n en que dos de los cuatro nodos (nodo 3 y 4) usan la interconexión del clúster para acceder al almacenamiento. Esta configuración puede ampliarse para que incluya nodos adicionales que no tengan acceso directo al almacenamiento compartido.

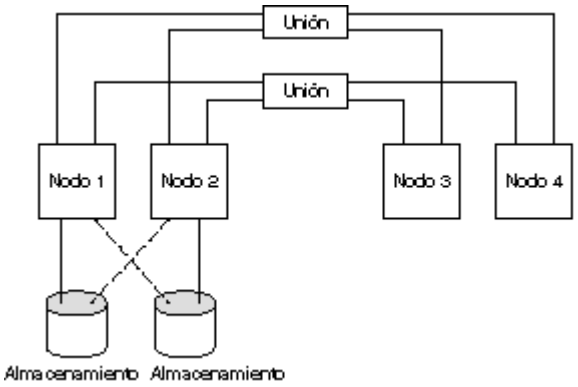


FIGURA 2–4 Topología de par+n

Topología n+1 (estrella)

Las topologías n+1 incluyen varios nodos primarios y uno secundario que no se deben configurar de forma idéntica. Los nodos primarios proporcionan de forma activa servicios de aplicación. El nodo secundario no debe estar desocupado mientras se espera que uno principal falle.

El nodo secundario es el único de la configuración que está conectado físicamente a todos las unidades de almacenamientos multisistema .

Si se produce un fallo en un nodo primario, Sun Cluster resuelve el error transfiriendo los recursos al secundario, dónde éstos funcionan hasta que se conmutan de nuevo (de forma automática o manual) al nodo primario.

El secundario siempre debe tener suficiente capacidad sobrante de CPU para manejar la carga si uno de los primarios falla.

La figura siguiente ilustra la configuración n+1.

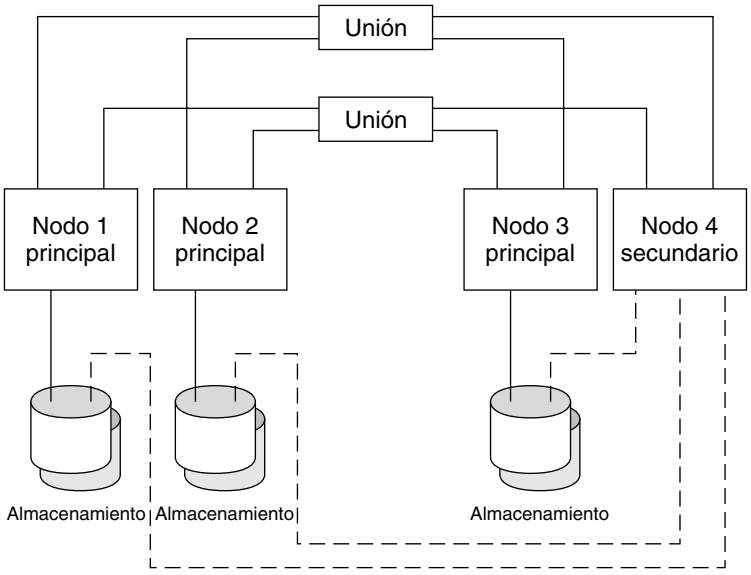


FIGURA 2–5 Topología n+1

Topología n*n (escalable)

Las topologías n*n permiten a todos los dispositivos de almacenaminto compartido del clúster conectarse con todos los nodos del clúster. Esta topología permite que las aplicaciones de alta disponibilidad se recuperen de fallos de un nodo a otro sin que se produzca una degradadación en el servicio. Cuando se produce una recuperación de fallos, el nodo nuevo puede acceder al dispositivo de almacenamiento usando una ruta local en lugar de la interconexión privada.

La figura siguiente ilustra la configuración n*n.

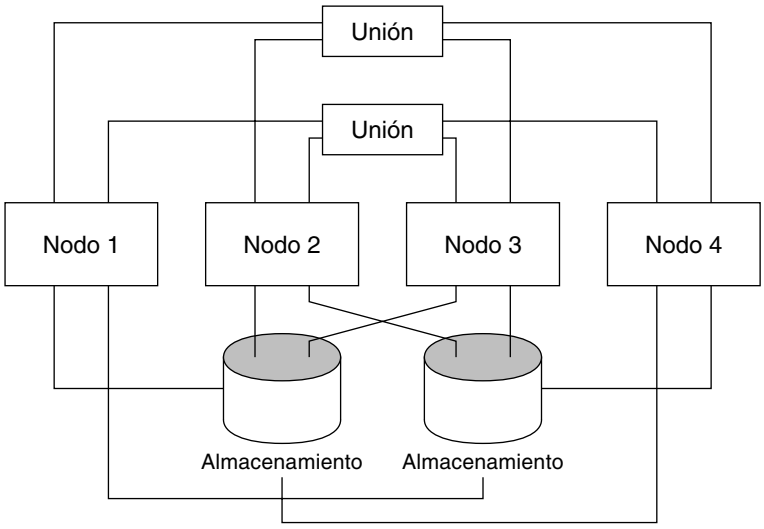


FIGURA 2–6 Topología n*n

CAPÍTULO 3

Conceptos clave de la administración y desarrollo de las aplicaciones

Este capítulo describe los conceptos clave relacionados con los componentes de hardware de una configuración de sistema de SunPlex. Se tratan los siguientes temas:

Administración del clúster y desarrollo de aplicaciones

Esta información va dirigida principalmente a administradores del sistema y desarrolladores de aplicaciones que utilicen API y SDK de SunPlex. Asimismo, a los administradores del sistema del clúster esta información les puede resultar útil para instalar, configurar y administrar el software del clúster. Los desarrolladores de aplicaciones pueden usar la información para entender el entorno del clúster en el que estarán trabajando.

La figura siguiente muestra una vista detallada que correlaciona los conceptos de la administración del clúster con su arquitectura.

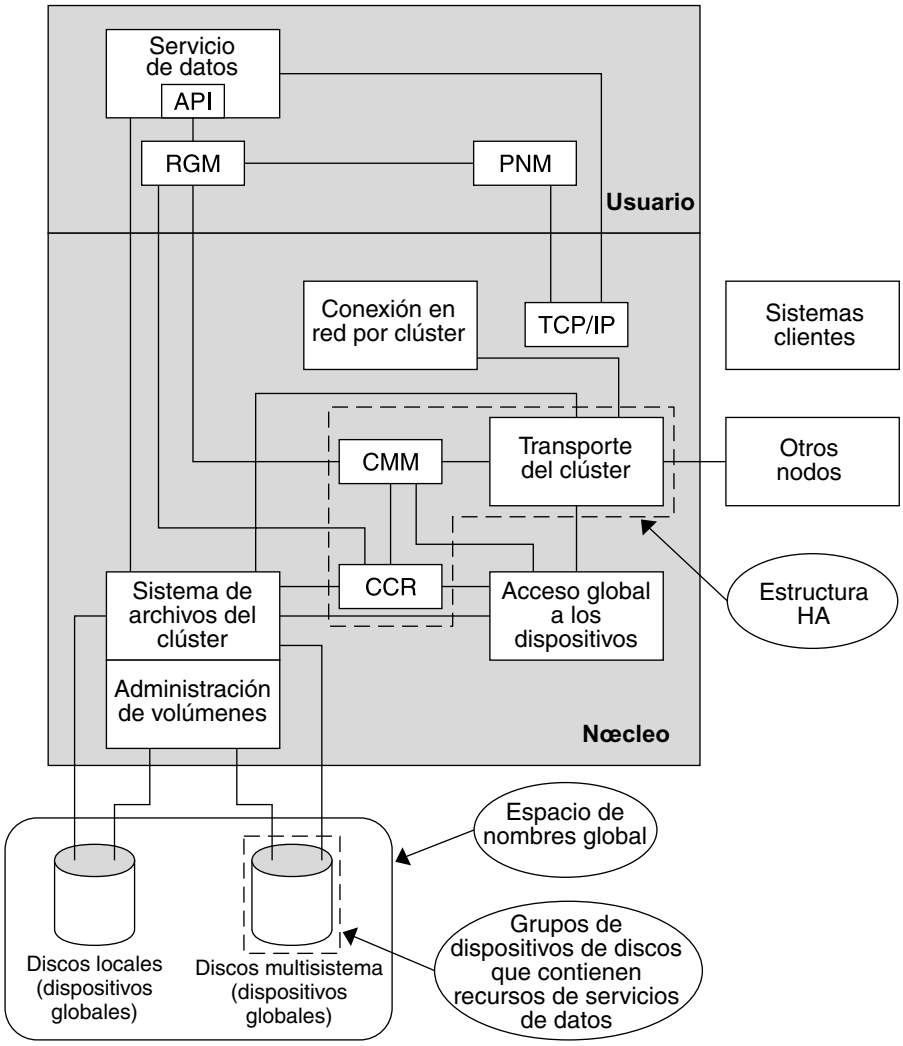


FIGURA 3–1 Arquitectura del software de Sun Cluster

Interfaces administrativas

Se puede elegir como instalar, configurar y administrar el sistema SunPlex desde varias interfaces de usuario, así como realizar tareas de administración del sistema a través de la interfaz gráfica de usuario (GUI) de SunPlex Manager o de la interfaz de

línea de comandos documentada. Además de la interfaz de línea de comandos hay utilidades como `scinstall` y `scsetup` que simplifican la instalación seleccionada y las tareas de configuración. El sistema SunPlex también tiene un módulo que se ejecuta como parte de Sun Management Center e incluye una GUI para algunas tareas del clúster. Consulte el capítulo de introducción de *Sun Cluster 3.1: Guía de administración del sistema* para obtener descripciones completas de las interfaces administrativas.

Hora del clúster

La hora en todos los nodos del clúster debe estar sincronizada. Que esta sincronización se realice con una fuente externa no es importante para el funcionamiento del clúster. El sistema SunPlex emplea Network Time Protocol (NTP) para sincronizar los relojes de los distintos nodos.

En general, una diferencia en el reloj del sistema de una fracción de segundo no causa problemas. Sin embargo, si ejecuta `date(1)`, `rdate(1M)` o `xntpdate(1M)` (interactivamente o desde secuencias `cron`) en un clúster activo, puede forzar un cambio de hora mucho mayor que el de una fracción de segundo para sincronizar el reloj del sistema con el origen de la hora, lo que podría causar problemas con la marca de hora de modificación de archivos o confundir al servicio NTP.

Cuando se instala el sistema operativo Solaris en todos los nodos del clúster, se tiene la oportunidad de cambiar la fecha y hora predeterminadas para el nodo. En general, puede aceptar el valor predeterminado sugerido.

Cuando se instala el software de Sun Cluster mediante `scinstall(1M)`, uno de los pasos del proceso es configurar NTP para el clúster. El software de Sun Cluster incluye un archivo de plantilla, `ntp.cluster` (consulte `/etc/inet/ntp.cluster` en un nodo del clúster instalado), que establece una relación entre iguales entre todos los nodos del clúster, de los cuales uno de ellos es el “preferido”. Los nodos se identifican por su nombre de sistema privado y la sincronización de la hora se produce entre la interconexión del clúster. Las instrucciones sobre cómo configurar el clúster para NTP se incluyen en *Sun Cluster 3.1: Guía de instalación del software*.

También puede configurar uno o más servidores NTP externos al clúster y cambiar el archivo `ntp.conf` para reflejar esta configuración.

Durante el funcionamiento normal, nunca debería ser necesario ajustar la hora en el clúster. Ahora bien, si la hora se estableció incorrectamente al instalar el sistema operativo Solaris y desea cambiarla, puede consultar el procedimiento para hacerlo en *Sun Cluster 3.1: Guía de administración del sistema*.

Estructura de alta disponibilidad (HA)

El sistema SunPlex hace que todos los componentes de la "ruta de acceso" entre usuarios y datos esté muy disponible, incluyendo las interfaces de red, las aplicaciones en sí, el sistema de archivos y los discos multisistema. En general, un componente del clúster está muy disponible si sobrevive a cualquier fallo (de software o hardware) que se produzca en el sistema.

La tabla siguiente muestra los tipos de fallos de componentes de SunPlex (tanto de hardware como de software) y los tipos de recuperación que incorpora la estructura de alta disponibilidad.

TABLA 3–1 Niveles de detección de fallos y recuperación en SunPlex

Componente del clúster fallido	Recuperación de software	Recuperación de hardware
Servicio de datos	API HA , estructura HA	N/D
Adaptador de red pública	Ruta múltiple de red IP	Varias tarjetas adaptadoras de red pública
Sistema de archivos del clúster	Réplicas primaria y secundaria	Discos multisistema
Disco multisistema duplicado	Gestión de volúmenes (Gestor de volúmenes de Solaris y VERITAS Volume Manager)	RAID-5 por hardware (por ejemplo: Sun StorEdge™ A3x00)
Dispositivo global	Réplicas primaria y secundaria	Varias rutas de acceso al dispositivo, uniones de transporte al clúster
Red privada	Software de transporte HA	Varias redes privadas independientes del hardware
Nodo	CMM, controlador de recuperación rápida	Varios nodos

La estructura de alta disponibilidad del software de Sun Cluster detecta el fallo en un nodo rápidamente y crea un servidor equivalente nuevo para los recursos de la estructura en uno de los nodos restantes del clúster. En ningún momento se interrumpe completamente la disponibilidad de los recursos de la estructura. Los que no se hayan visto afectados por el nodo que haya fallado están completamente disponibles durante la recuperación. Además, los recursos de la estructura del nodo fallido vuelven a estar disponibles en cuanto se recuperan. Un recurso de la estructura recuperado no tiene que esperar a que todos los demás recursos de la estructura completen su recuperación.

La mayoría de recursos de la estructura de alta disponibilidad se recuperan de manera transparente para las aplicaciones (servicios de datos) usando el recurso. Las semánticas del acceso a los recursos de la estructura se conservan perfectamente entre

los fallos de los nodos. Las aplicaciones simplemente no pueden advertir que el servidor del recurso de la estructura se ha trasladado a otro nodo. El fallo de un único nodo es completamente transparente para los programas desde el resto de los nodos que usan archivos, dispositivos y volúmenes de disco conectados a este nodo, siempre que exista una ruta de acceso alternativa de hardware a los discos desde otro nodo. Un ejemplo es el uso de discos multisistema que tengan puertos con varios nodos.

Supervisor de pertenencia al clúster (CMM)

El Supervisor de pertenencia al clúster es un conjunto distribuido de agentes, uno por miembro del clúster, que intercambian mensajes por la interconexión del clúster para:

- Forzar una vista de miembros uniforme en todos los nodos (quórum)
- Activar la reconfiguración sincronizada en respuesta a cambios en la pertenencia, mediante retrollamadas registradas
- Manejar el particionamiento del clúster (esquizofrenia, amnesia)
- Asegurar la completa conectividad entre todos los miembros del clúster

A diferencia de anteriores versiones del software Sun Cluster, CMM se ejecuta completamente en el núcleo.

Pertenencia al clúster

La función principal de CMM es establecer acuerdos a nivel del clúster sobre el conjunto de nodos que participan en éste en todo momento. Esta limitación se denomina *pertenencia al clúster*.

Para determinar la pertenencia al clúster y, por tanto, asegurar la integridad de los datos, CMM:

- Registra los cambios en la pertenencia al clúster, como la incorporación o el cese de un nodo en el clúster
- Se asegura que los nodos “defectuosos” abandonen el clúster
- Se asegura que los nodos “defectuosos” permanezcan fuera del clúster hasta que éste se repare
- Evita que el clúster se particione en subgrupos de nodos

Consulte «Quórum y dispositivos del quórum» en la página 52 para obtener más información sobre cómo se protege el clúster de particionarse en varios clústers independientes.

Reconfiguración del supervisor de pertenencia al clúster

Para asegurarse de que los datos permanezcan incorruptos, 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 (aplicaciones) en respuesta a un fallo.

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

Después de detectar un cambio en la composición del clúster, CMM lleva a cabo una configuración sincronizada del clúster en que los recursos de éste podrían redistribuirse de acuerdo con la nueva composición.

Mecanismo de recuperación rápida

Si CMM detecta un problema crítico en un nodo, envía una señal a través de la estructura del clúster para forzar su apagado (pánico) y así retirar su pertenencia al clúster. El mecanismo por el que ello ocurre se denomina *recuperación rápida*. Éste obliga a un nodo a apagarse de dos formas.

- Si un nodo abandona el clúster y después intenta crear uno nuevo sin tener quórum, queda "encerrado" y se le impide acceder a discos compartidos. Consulte «Aislamiento de fallos» en la página 55 para obtener detalles sobre este uso de la recuperación rápida.
- Si uno o más daemons específicos del clúster dejan de existir (`clexecd`, `rpc.pmfd`, `rgmd` o `rpc.ed`) CMM detecta el fallo y el nodo entra en pánico.

Cuando la desaparición de un daemon del clúster hace que un nodo entre en pánico, aparecerá un mensaje parecido a éste en la consola de ese nodo.

```
panic[cpu0]/thread=40e60: Failfast: Aborting because "pmfd" died 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
```

Después de la condición de pánico, el nodo podría rearrancar e intentar volverse a unir al clúster o permanecer en el indicador OpenBoot™ PROM (OBP). La acción que se toma depende del valor del parámetro `auto-boot?` en la OBP.

Depósito de configuración del clúster (CCR)

El depósito de configuración del clúster (CCR) es una base de datos privada compartida a nivel del clúster para almacenar información perteneciente a la configuración y estado del clúster. CCR es una base de datos distribuida. Cada nodo mantiene una copia completa de la base de datos. CCR se asegura de que todos los nodos tengan una vista uniforme de la "extensión" del nodo. Para evitar que se corrompan datos, todos los nodos necesitan conocer el estado actual de los recursos del clúster.

CCR usa un algoritmo de puesta al día de dos fases para las actualizaciones: una actualización se tiene que completar satisfactoriamente en todos los miembros del clúster o de lo contrario se anula. CCR usa la interconexión del clúster para aplicar las actualizaciones distribuidas.

Precaución – CCR se compone de archivos de texto que nunca se deben editar manualmente, ya que cada archivo contiene un registro de suma de control para asegurar la coherencia entre los nodos. La actualización manual de los archivos de CCR provocaría que un nodo o todo el clúster dejaran de funcionar.

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.

Dispositivos globales

El sistema SunPlex usa *dispositivos globales* para ofrecer a todo el clúster un acceso de alta disponibilidad a cualquier dispositivo del clúster, desde cualquier nodo sin que importe dónde esté la conexión física del dispositivo. En general, si un nodo falla mientras se ofrece acceso a un dispositivo global, el software de Sun Cluster descubre automáticamente otra ruta de acceso al dispositivo y redirige el acceso a la misma. Los dispositivos globales de SunPlex pueden ser discos, CD-ROM y cintas. Sin embargo, los discos son los únicos dispositivos globales multipuerto 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 automáticamente ID exclusivas a cada disco, CD-ROM y unidad de cinta del clúster, lo que permite el acceso uniforme a todos los dispositivos desde cualquier nodo del clúster. El espacio de nombres de dispositivos global se conserva en el directorio `/dev/global`. Para obtener más información, véase «Espacio de nombres global» en la página 44.

Los dispositivos globales multipuerto ofrecen más de una ruta de acceso al dispositivo. En el caso de discos multisistema, debido a que forman parte de un grupo de dispositivos de disco alojado por más de un nodo, eso los convierte en discos de alta disponibilidad.

ID de dispositivo (DID)

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

El pseudocontrolador de ID de dispositivo (DID) forma parte integral de la prestación de acceso global a los dispositivos del clúster. El controlador de DID examina todos los nodos del clúster y crea una lista de dispositivos de disco única, asignándole a cada uno un número mayor y menor exclusivos que es idéntico en todos los nodos del clúster. El acceso a los dispositivos globales se realiza usando el ID de dispositivo único asignado por el controlador DID en lugar de usar los tradicionales ID de dispositivo de Solaris, como `c0t0d0` para un disco.

Este enfoque fuerza a que todas las aplicaciones que acceden a discos (como un gestor de volúmenes o aplicaciones que usen dispositivos a bajo nivel) utilicen una ruta de acceso uniforme en todo el clúster. Esta uniformidad es especialmente importante en discos multisistema, debido a que los números mayor y menor de cada dispositivo pueden variar entre distintos nodos, cambiando así también las convenciones de asignación de nombres de dispositivos de Solaris. Por ejemplo, el nodo1 podría ver un disco multisistema como `c1t2d0` y el nodo2 podría ver ese mismo disco de forma completamente distinta, como `c3t2d0`. El controlador DID asigna un nombre global, como d10, que los nodos usarán en su lugar dando a cada uno una correlación uniforme con el disco multisistema.

Los ID de dispositivo se administran con los comandos `scdidadm(1M)` y `scgdevs(1M)`. Consulte las páginas de comando man respectivas para obtener mas información.

Grupos de dispositivos de disco

En el sistema SunPlex, todos los discos multisistema deben estar bajo el control del software Sun Cluster. En primer lugar en los discos multisistema se crean grupos de discos del gestor de volúmenes, bien conjuntos de discos del Gestor de volúmenes de Solaris, bien grupos de discos de VERITAS Volume Manager. A continuación, se registran los grupos de discos del gestor de discos como *grupos de dispositivos de discos* que forman un tipo de dispositivo global. Además, el software Sun Cluster crea automáticamente un grupo de dispositivos de bajo nivel para cada dispositivo de disco y de cinta del clúster. Sin embargo, estos grupos de dispositivos del clúster permanecen en estado fuera de línea hasta que se accede a ellos como dispositivos globales.

El registro proporciona a SunPlex información del sistema sobre los nodos y sus rutas de acceso a grupos de disco del gestor de volúmenes. En este punto, los grupos de discos del gestor de volúmenes se convierten en accesibles globalmente dentro del clúster. Si hay más de un nodo que pueda escribir (controlar) un grupo de dispositivos de disco, los datos almacenados en éste se consideran de alta disponibilidad. Estos grupos pueden usarse para alojar sistemas de archivos del clúster.

Nota – Los grupos de dispositivos de disco son independientes de los grupos de recursos. Un nodo puede controlar un grupo de recursos (que represente un grupo de procesos de servicio de datos) mientras otro puede controlar los grupos de discos a los que están accediendo los servicios de datos. Sin embargo, la mejor práctica es mantener en el mismo nodo el grupo de dispositivos de disco que almacena los datos de una aplicación determinada y el grupo que contiene sus recursos (el daemon de la aplicación). Consulte el capítulo de resumen de *Sun Cluster 3.1 Data Services Installation and Configuration Guide* para obtener más información sobre la asociación entre grupos de dispositivos de disco y grupos de recursos.

Con un grupo de dispositivos de disco, el grupo de disco del gestor de volúmenes se convierte en "global" porque proporciona soporte de ruta múltiple a los discos que incluye. Cada nodo del clúster conectado físicamente a los discos multisistema proporciona una ruta de acceso al grupo de dispositivos de disco.

Recuperación de fallos del grupo de dispositivos de disco

Debido a que un alojamiento de disco está conectado a más de un nodo, todos los grupos de dispositivos de disco de ese alojamiento estarán disponibles a través de una ruta de acceso alternativa si falla el nodo que esté controlando en ese momento el grupo de dispositivos. El fallo del nodo que controla el grupo de dispositivos no afecta al acceso al grupo de dispositivos excepto por el tiempo que se tarda en realizar la recuperación y las comprobaciones de integridad. Durante ese tiempo, todas las peticiones se bloquean (de forma transparente para la aplicación) hasta que el sistema vuelve a hacer que el grupo de dispositivos esté disponible.

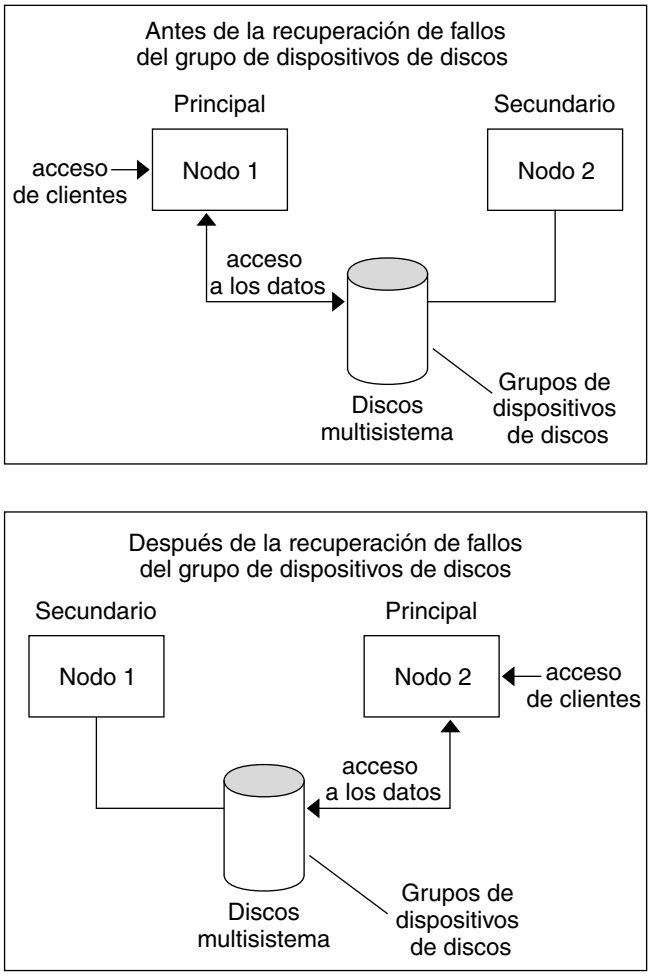


FIGURA 3–2 Recuperación de fallos en grupos de dispositivos de disco

Grupos de dispositivos de disco multipuerto

Este apartado describe las propiedades del grupo de dispositivo de disco que permiten equilibrar el rendimiento y la disponibilidad en una configuración de disco multipuerto. El software Sun Cluster proporciona dos propiedades que se usan para ajustar una configuración de disco multipuerto: `preferenced` y `numsecondaries`. Con la propiedad `preferenced` se puede controlar el orden en el que los nodos intentan asumir el control cuando se produce una recuperación de fallo. La propiedad `numsecondaries` se utiliza para establecer un número deseado de nodos secundarios para un grupo de dispositivos.

Un servicio de alta disponibilidad se considera caído cuando el nodo primario cae y no hay otros secundarios que puedan promocionar a los primarios. Si se produce la recuperación de fallos y la propiedad `preferenced` es `true`, los nodos siguen el orden de la lista para que se seleccione un secundario. La lista de nodos configurada define el orden en que que éstos intentarán asumir el control primario o transicionar de redundante a secundario. Puede cambiar dinámicamente la preferencia de un servicio de dispositivo mediante la utilidad `scsetup`(1M). La preferencia que está asociada con proveedores de servicio dependientes, por ejemplo un sistema de archivos global, será la del servicio del dispositivo.

El nodo primario comprueba los secundarios durante el funcionamiento normal. En una configuración de disco multipuerto, comprobar todos los nodos secundarios produce una degradación en el rendimiento del clúster y una sobrecarga en la memoria. El soporte para nodos redundantes se ha implementado para minimizar la degradación en el rendimiento y la sobrecarga de memoria que produce el proceso de comprobación. De forma predeterminada el grupo de dispositivos de disco tendrá uno primario y uno secundario. El resto de nodos de proveedor disponibles estarán en línea en el estado redundante. Si se produce una recuperación de fallos, el secundario se convertirá en el primario y el nodo con más prioridad de la lista se convertirá en el secundario.

El número de nodos secundarios que se desee se puede establecer en cualquier número entero entre uno y el número de nodos proveedores no primarios operativos del grupo de dispositivos.

Nota – Si se está utilizando Solaris Volume Manager, se debe crear el grupo de dispositivos de disco antes de establecer la propiedad `numsecondaries` a un número distinto del predeterminado.

De manera predeterminada el número deseado de secundarios para servicios de dispositivos es de uno. El número real de proveedores secundarios que mantiene la estructura de réplica es la deseada, a menos que el número de proveedores no primarios en funcionamiento sea inferior al deseado. Si está añadiendo o quitando nodos de la configuración, deberá hacer cambios en la propiedad `numsecondaries` y revisar bien la lista de nodos. Si se mantiene la lista de nodos y el número deseado de secundarios se evitarán conflictos entre el número de secundarios configurado y el número real permitido por la estructura. Para gestionar la incorporación y eliminación de nodos de la configuración, use el comando `scconf`(1M) para grupos de discos VxVM o el comando `metaset`(1M) para grupos de dispositivos del Gestor de volúmenes de Solaris junto con los valores de configuración `preferenced` y `numsecondaries`. Consulte “Administering Global Devices and Cluster File Systems” in *Sun Cluster 3.1 System Administration Guide* para obtener información de procedimientos sobre cómo modificar las propiedades de los grupos de dispositivos de disco.

Espacio de nombres global

El mecanismo de software de Sun Cluster que habilita los dispositivos globales es el *espacio de nombres global* que incluye la jerarquía `/dev/global/` así como los espacios de nombres del gestor de volúmenes. El espacio de nombres global representa los discos multisistema y locales (y cualquier otro dispositivo del clúster, como CD-ROM y cintas) e incluye varias rutas de acceso de recuperación de fallos a los discos multisistema. Todos los nodos conectados físicamente a discos multisistema incluyen una ruta de acceso al almacenamiento válida para todos los nodos del clúster.

Normalmente los espacios de nombres del gestor de volúmenes residen en los directorios `/dev/md/`*conjunto-discos*`/dsk` (y `rdsk`) para el Gestor de volúmenes de Solaris y en los directorios `/dev/vx/dsk/`*grupo-disco* y `/dev/vx/rdsk/`*grupo-disco* para VxVM. Estos espacios de nombres se componen de directorios para cada conjunto de discos del Gestor de volúmenes de Solaris y cada grupo de discos de VxVM importados a través del clúster, respectivamente, cada uno de los cuales aloja un nodo de dispositivo para cada metadispositivo o volumen de ese conjunto o grupo de discos.

En el sistema SunPlex todos los nodos de dispositivos del espacio de nombres del gestor de volúmenes local se sustituye por un enlace simbólico con un nodo de dispositivo del sistema de archivos `/global/.devices/node@`*IDnodo*, donde *IDnodo* es un número entero que representa los nodos del clúster. El software de Sun Cluster sigue presentando los dispositivos del gestor de volúmenes, como enlaces simbólicos, también en sus ubicaciones estándar. Tanto el espacio de nombres global como el estándar del gestor de volúmenes están disponibles desde cualquier nodo del clúster.

Entre las ventajas del espacio de nombres global, se cuentan:

- Cada nodo permanece bastante independiente, con pocos cambios en el modelo de administración de dispositivos.
- Los dispositivos puede convertirse en globales de forma selectiva.
- Los generadores de enlaces de terceros siguen funcionando.
- Dado un nombre de dispositivo local, se ofrece una correlación simple para obtener un nombre global.

Ejemplo de espacios de nombres locales y globales

La tabla siguiente muestra las correlaciones entre los espacios de nombres local y global para un disco multisistema, `c0t0d0s0`.

TABLA 3–2 Correlaciones de espacios de nombres local y global

Componente/Ruta de acceso	Espacio de nombres de nodo local	Espacio de nombres global
Nombre lógico de Solaris	`/dev/dsk/c0t0d0s0`	`/global/.devices/node@`*IDnodo*`/dev/dsk/c0t0d0s0`
Nombre DID	`/dev/did/dsk/d0s0`	`/global/.devices/node@`*IDnodo*`/dev/did/dsk/d0s0`
Gestor de volúmenes de Solaris	`/dev/md/`*conjunto-discos*`/dsk/d0`	`/global/.devices/node@`*IDnodo*`/dev/md/`*diskset*`/dsk/d0`
VERITAS Volume Manager	`/dev/vx/dsk/`*grupo-discos*`/v0`	`/global/.devices/node@`*IDnodo*`/dev/vx/dsk/`*grupo-discos*`/v0`

El espacio de nombres global se genera automáticamente durante la instalación y se actualiza en cada arranque de reconfiguración; también se puede generar mediante el comando `scgdevs(1M)`.

Sistemas de archivos del clúster

Un sistema de archivos del clúster es un delegado entre el núcleo de un nodo y el sistema de archivos subyacente y el gestor de volúmenes que se ejecuta en un nodo que tiene una conexión física con uno o varios discos.

Los sistemas de archivos del clúster dependen de dispositivos globales (discos, cintas, CD-ROM) con conexiones físicas a uno o más nodos. A los dispositivos globales se puede acceder desde cualquier nodo del clúster a través del mismo nombre de archivo (por ejemplo, `/dev/global/`) aunque el nodo no tenga una conexión física con el dispositivo de almacenamiento. Los dispositivos globales se pueden usar igual que los normales, es decir, se puede crear un sistema de archivos en ellos usando `newfs` o `mkfs`.

Se puede montar un sistema de archivos en un dispositivo global con `mount -g` (globalmente) o con `mount` (localmente).

Los programas pueden acceder a los archivos del sistema de archivos del clúster desde cualquier nodo de éste empleando el mismo nombre de archivo (por ejemplo, `/global/foo`).

Los sistemas de archivos del clúster se montan en todos los miembros del clúster, pero no puede montarse en un subconjunto de miembros del clúster.

Los sistemas de archivo clúster no son de tipo diferenciado. Es decir, los clientes ven el sistema de archivos subyacente (por ejemplo, UFS).

Uso de los sistemas de archivos del clúster

En el sistema SunPlex todos los discos multisistema se sitúan en grupos de dispositivos de disco que pueden ser conjuntos de discos del Gestor de volúmenes de Solaris, grupos de discos de VxVM o discos individuales que no están bajo el control de ningún gestor de volúmenes basado en software.

Para que un sistema de archivos del clúster sea de alta disponibilidad, el almacenamiento en disco subyacente debe estar conectado a más de un nodo. Por consiguiente, un sistema de archivos local (aquél que está almacenado el disco local de un nodo) que se convierte en sistema de archivos del clúster no es de alta disponibilidad.

Al igual que ocurre con los sistemas de archivo locales, los sistemas de archivo clúster se pueden montar de dos formas:

- **Manualmente**: con el comando `mount` y las opciones de montaje `-g` u `-o global` para montar el sistema de archivos del clúster desde la línea de comandos, por ejemplo:

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

- **Automáticamente**: creando una entrada en el archivo `/etc/vfstab` con una opción de montaje `global` para montar el sistema de archivos del clúster desde el arranque. Después puede crear un punto de montaje en el directorio `/global` de todos los nodos. Éste es una ubicación recomendada, no un requisito. La siguiente es una línea de ejemplo para un sistema de archivos del clúster del archivo `/etc/vfstab`:

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

Nota – Aunque el software de Sun Cluster no impone ninguna política de asignación de nombres a los sistemas de archivos del clúster, puede facilitarse la administración creando un punto de montaje para todos los sistemas de archivos del clúster en el mismo directorio, como `/global/`*grupo-dispositivo-disco*. Para obtener mas información, consulte *Sun Cluster 3.1: Guía de instalación del software* y *Sun Cluster 3.1: Guía de administración del sistema*.

Prestaciones del sistema de archivos del clúster

El sistema de archivos del clúster dispone de las prestaciones siguientes:

- Las ubicaciones de los accesos de archivo son transparentes. Un proceso puede abrir un archivo situado en cualquier parte del sistema y los procesos de todos los nodos pueden usar el mismo nombre de ruta para situar un archivo.

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

- 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 la funcionalidad de bloqueo de archivos a través de las interfaces `fcntl(2)`. Las aplicaciones que se ejecutan en varios nodos del clúster pueden sincronizar el acceso a los datos mediante el bloqueo de archivo condicional en un 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 mientras siga 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 convierten en global cualquier sistema de archivos en disco admitido.

Tipo de recurso `HAStoragePlus`

El tipo de recurso `HAStoragePlus` está diseñado para convertir en altamente disponibles configuraciones de sistemas de archivos no globales como UFS y VxFS, permite integrar el sistema de archivos local en el entorno Sun Cluster, convertir aquél en altamente disponible y proporcionar prestaciones del sistema de archivos adicionales, como comprobaciones, montajes y desmontajes forzados que permiten a Sun Cluster recuperarse pasando a sistemas de archivos locales. Para la recuperación de fallos el sistema de archivos local debe residir en grupos de discos globales que tengan habilitados conmutadores de afinidad.

Consulte los capítulos de servicios de datos individuales en *Data Services Installation and Configuration Guide* o el capítulo 14 "Enabling Highly Available Local File Systems" de "Administering Data Services Resources" para obtener información sobre cómo usar el tipo de recurso `HAStoragePlus`.

Éste también se puede usar para sincronizar el inicio de recursos y grupos de dispositivos de disco de los que dependen los recursos. Para obtener más información, consulte «Recursos: definición, grupos y tipos» en la página 69.

Opción de montaje syncdir

La opción de montaje `syncdir` puede utilizarse en sistemas de archivos del clúster que usen UFS como sistema de archivo subyacente. Sin embargo, existe una mejora de rendimiento significativa si no se especifica `syncdir`. Si se especifica `syncdir`, se garantiza que las escrituras sean compatibles con POSIX. Si no se especifica, tendrá el mismo comportamiento que se ve con sistemas de archivos NFS. Por ejemplo, en algunos casos sin `syncdir`, no se descubriría una condición de falta de espacio hasta que se cerrara el archivo. Con `syncdir` (y el comportamiento POSIX), la condición de falta de espacio se descubriría durante la operación de escritura. Los casos en los que se pueden tener problemas si no se especifica `syncdir` son raros, por lo que se recomienda no especificarlo y aprovechar la ventaja del mejor rendimiento.

VxFS no dispone de una opción de montaje equivalente a `syncdir` para UFS. El comportamiento de VxFS es el mismo que para UFS cuando no se especifica la opción de montaje `syncdir`.

Consulte «FAQ sobre sistemas de archivos» en la página 88 para obtener preguntas frecuentes sobre los dispositivos globales y los sistemas de archivos del clúster.

Supervisión de las rutas del disco

La versión actual del software Sun Cluster admite la supervisión de las rutas del disco (DPM). Este apartado incluye información conceptual sobre DPM, el daemon DPM y las herramientas de administración que se pueden usar para supervisar las rutas del disco. Consulte *Sun Cluster 3.1 9/03 System Administration Guide* para obtener información de los procedimientos para supervisar, dejar de supervisar y comprobar el estado de las rutas del disco.

Nota – DPM no se admite en nodos que ejecuten versiones anteriores al software Sun Cluster 3.1 5/03. No utilice comandos de DPM durante una modernización. Una vez finalizada la modernización en todos los nodos, éstos deben estar en línea para poder utilizar los comandos de DPM.

Visión general

DPM mejora la disponibilidad general de la recuperación de fallos y la conmutación por supervisión de la disponibilidad de la ruta de acceso a disco secundaria. Utilice el comando `scdpm` para verificar la disponibilidad de la ruta del disco que utiliza un recurso antes de conmutarlo. Las opciones que se incluyen con el comando `scdpm` permiten supervisar las rutas de acceso a discos hacia un nodo individual o hacia todos los nodos del clúster. Consulte la página de comando man `scdpm`(1M) para obtener más información sobre las opciones de la línea de comandos.

Los componentes DPM se instalan a partir del paquete `SUNWscu`. Éste lo instala el procedimiento de instalación estándar de Sun Cluster. Consulte la página de comando man `scinstall`(1M) para obtener detalles sobre la instalación de la interfaz. La tabla siguiente describe la ubicación predeterminada de los componentes DPM.

Ubicación	Componente
Daemon	`/usr/cluster/lib/sc/scdpmd`
Interfaz de línea de comandos	`/usr/cluster/bin/scdpm`
Bibliotecas compartidas	`/user/cluster/lib/libscdpm.so`
Archivo de estado de daemon (creado en tiempo de ejecución)	`/var/run/cluster/scdpm.status`

En cada nodo se ejecuta un daemon DPM de subproceso múltiple. El daemon DPM (`scdpmd`) lo inicia la secuencia `rc.d` cuando arrancan los nodos. Si surge algún problema, `pmfd` gestiona el daemon y lo reinicia automáticamente. La lista siguiente describe cómo funciona `scdpmd` en el arranque inicial.

Nota – En el arranque, el estado de cada ruta del disco se inicializa a `UNKNOWN`.

---

1. El daemon DPM recoge información de rutas del disco y nombres de nodo del archivo de estado anterior o de la base de datos CCR. Para obtener más información sobre CCR, consulte «Depósito de configuración del clúster (CCR)» en la página 38. Cuando se inicia el daemon DPM, éste puede forzarse para que lea la lista de discos supervisados a partir de un nombre de archivo especificado.
2. El daemon DPM inicializa la interfaz de comunicaciones para que responda a solicitudes de componentes que son externos al mismo, como la interfaz de línea de comandos.
3. El daemon DPM realiza un ping a cada ruta del disco de la lista supervisada cada 10 minutos mediante comandos de tipo `scsi_inquiry`. Todas las entradas están bloqueadas para evitar que la interfaz de comunicaciones acceda al contenido de una entrada que esté siendo modificada.
4. El daemon DPM envía una notificación a Sun Cluster Event Framework y registra el estado nuevo de la ruta a través del mecanismo UNIX `syslogd`(1M).

Nota – Todos los errores relacionados con el daemon se tratan en `pmfd`(1M). Todas las funciones de la API devuelven `0` cuando tienen éxito y `-1` cuando se produce algún error.

El daemon DPM supervisa la disponibilidad de la ruta lógica que es visible a través de controladores de ruta múltiple como MPxIO, HDLM y PowerPath. Las rutas de acceso físicas individuales gestionadas por estos controladores no están supervisadas porque el controlador de ruta múltiple oculta los fallos individuales al daemon DPM.

Supervisión de las rutas del disco

Este apartado describe dos métodos para supervisar las rutas del disco en el clúster. El primer método lo ofrece el comando `scdpm`. Utilice este comando para supervisar, dejar de supervisar o mostrar el estado de las rutas del disco del clúster. Este comando también resulta útil para imprimir la lista de discos fallidos y supervisar rutas de los discos a partir de un archivo.

El segundo método para supervisar las rutas de los discos en el clúster lo ofrece la interfaz gráfica de usuario (GUI) de SunPlex Manager. SunPlex Manager incluye una vista topográfica de las rutas del disco supervisadas del clúster. La vista se actualiza cada 10 minutos para incluir información sobre el número de pings que han fallado. Para administrar rutas del disco use la información que proporciona la GUI de SunPlex Manager junto con el comando `scdpm`(1M). Consulte "Administering Sun Cluster With the Graphical User Interfaces" in *Sun Cluster 3.1 9/03 System Administration Guide* para obtener información sobre SunPlex Manager.

Uso del comando `scdpm` para supervisar las rutas del disco

El comando `scdpm`(1M) proporciona a DPM comandos de administración que permiten realizar las tareas siguientes:

- Supervisar una ruta del disco nueva
- Dejar de supervisar una ruta del disco
- Volver a leer los datos de configuración de la base de datos CCR
- Leer los discos para supervisar o dejar de hacerlo a partir de un archivo especificado
- Informar del estado de una o todas las rutas de acceso de disco del clúster
- Imprimir todas las rutas de acceso de disco que son accesibles desde un nodo

Emitir el comando `scdpm`(1M) con el argumento ruta-disco desde cualquier nodo para llevar a cabo tareas de administración de DPM en el clúster. El argumento de ruta-disco siempre está compuesto por un nombre de nodo y uno de disco. El primer nodo no es necesario y, en caso de no especificarlo, toma el valor `all`. La tabla siguiente describe las convenciones de asignación de nombres para la ruta del disco.

Nota – Se recomienda utilizar nombres de ruta del disco globales, ya que son coherentes dentro de todo el clúster. Los nombres de las rutas del disco UNIX no son coherentes en todo el clúster. La ruta del disco UNIX correspondiente a un disco determinado puede diferir en distintos nodos del clúster. La ruta podría ser `c1t0d0` en un nodo y `c2t0d0` en otro. Si utiliza nombres de ruta del disco UNIX, utilice el comando `scdidadm -L` para asignar el nombre UNIX al nombre global antes de ejecutar comandos de DPM. Consulte la página de comando man `scdidadm`(1M).

TABLA 3–3 Ejemplos de nombres de rutas del disco

Tipo de nombre	Ejemplo de nombre de ruta de disco	Descripción
Ruta de disco global	`schost-1:/dev/did/dsk/d1`	Ruta de disco `d1` en el nodo `schost-1`
	`all:d1`	Ruta de disco `d1` en todos los nodos del clúster
Ruta del disco UNIX	`schost-1:/dev/rdsk/c0t0d0s0`	Ruta del disco `c0t0d0s0` en el nodo `schost-1`
	`schost-1:all`	Todas las rutas del nodo `schost-1`
Todas las rutas del disco	`all:all`	Todas las rutas del disco de todos los nodos del clúster

Uso de SunPlex Manager para supervisar las rutas del disco

SunPlex Manager permite realizar las siguientes tareas de administración DPM básicas:

- Supervisión de una ruta del disco
- Dejar de supervisar una ruta del disco
- Visualizar el estado de todas las rutas del disco del clúster.

Consulte la ayuda en línea de SunPlex Manager para obtener información de procedimientos para realizar una administración de la ruta del disco utilizando SunPlex Manager.

Quórum y dispositivos del quórum

Debido a que los nodos del clúster comparten datos y recursos, es importante no dividir nunca un clúster en particiones distintas que estén activas al mismo tiempo. CMM garantiza que como máximo haya un solo clúster en marcha en todo momento, incluso aunque la interconexión del clúster esté particionada.

Hay dos tipos de problemas que surgen a partir de las particiones de los clústers: esquizofrenia y amnesia. La esquizofrenia se produce cuando se pierde la interconexión del clúster entre nodos y éste se particiona en subclústers, cada uno los cuales cree que es la única partición. Es una consecuencia de problemas de comunicación entre los nodos del clúster. La amnesia se produce cuando el clúster se reinicia después de un apagado teniendo datos del clúster más antiguos que los del momento del apagado. Esto puede ocurrir si hay varias versiones de los datos de la estructura almacenadas en disco y se inicia una nueva copia del clúster cuando la más reciente no está disponible.

La esquizofrenia y la amnesia se pueden evitar dando a cada nodo un voto y obligando 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 votación mayoritaria funciona bien mientras haya más de dos nodos por clúster. En un clúster de dos nodos, la mayoría es dos. Si ese tipo del clúster resulta particionado es necesario un voto externo para que alguna de las dos particiones obtenga el quórum. Este voto externo lo proporciona un *dispositivo del quórum* que puede ser cualquier disco que compartan dos nodos. Los discos usados como dispositivos del quórum pueden contener datos de usuario.

La Tabla 3–4 describe cómo el software de Sun Cluster usa el quórum para evitar la esquizofrenia y la amnesia.

TABLA 3–4 Quórum del clúster y problemas de esquizofrenia y amnesia

Tipo de partición	Solución del quórum
Esquizofrenia	Sólo permite que la partición (sub-clúster) con mayoría de votos se ejecute como clúster (donde como máximo puede existir una partición con mayoría); cuando el nodo pierde la votación y no reúne el quórum necesario, éste entra en pánico
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)

El algoritmo de quórum funciona dinámicamente: a medida que los eventos del clúster accionan cálculos, los resultados de éstos pueden cambiar durante la vida del clúster.

Recuentos de votos del quórum

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 también pueden tener un recuento de votos de cero, por ejemplo cuando 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).

Los dispositivos del quórum se configuran durante la instalación del clúster o más adelante mediante los procedimientos que se describen en *Sun Cluster 3.1: Guía de administración del sistema*.

Nota – Los dispositivos del quórum sólo contribuyen al recuento de votos si al menos uno de los nodos al que está conectado es un miembro del clúster. Además, durante el arranque del clúster un dispositivo del quórum contribuye al recuento sólo si, al menos, uno de los nodos a los que está conectado en ese momento está arrancando y era miembro del clúster que ha arrancado más recientemente.

Configuraciones del quórum

Las configuraciones del quórum dependen del número de nodos del clúster:

- **Clústers de dos nodos**: para que se forme un clúster de dos nodos hacen falta dos votos de quórum que pueden provenir de los dos nodos del clúster o de un nodo y un dispositivo del quórum. Sin embargo, es necesario configurar un dispositivo del quórum en clústers de dos nodos para asegurarse que uno de ellos pueda continuar si el otro falla.
- **Clústers de más de dos nodos**: se debe especificar un dispositivo del quórum entre todos los pares de nodos que compartan acceso al alojamiento del almacenamiento de disco. Por ejemplo, supongamos que hay un clúster de tres nodos parecido al que se muestra en la Figura 3–3. En esta figura, el nodo A y el nodo B comparten acceso al mismo alojamiento de disco y el nodo B y el nodo C a otro distinto. Existiría pues un total de cinco votos de quórum, tres de los nodos y dos de los dispositivos del quórum compartidos entre los nodos. Un clúster necesita una mayoría de los votos de quórum para formarse.

  El software de Sun Cluster no obliga a especificar un dispositivo del quórum entre cada par de nodos que comparta acceso a un alojamiento de almacenamiento en disco. Sin embargo, puede proporcionar los votos de quórum necesarios para el caso en que una configuración N+1 degenere en un clúster de dos nodos y también

falle el nodo con acceso a ambos alojamientos de disco. Si ha configurado dispositivos del quórum entre todos los pares, el nodo restante podría seguir funcionando como clúster.

Consulte la Figura 3–3 para obtener ejemplos de estas configuraciones.

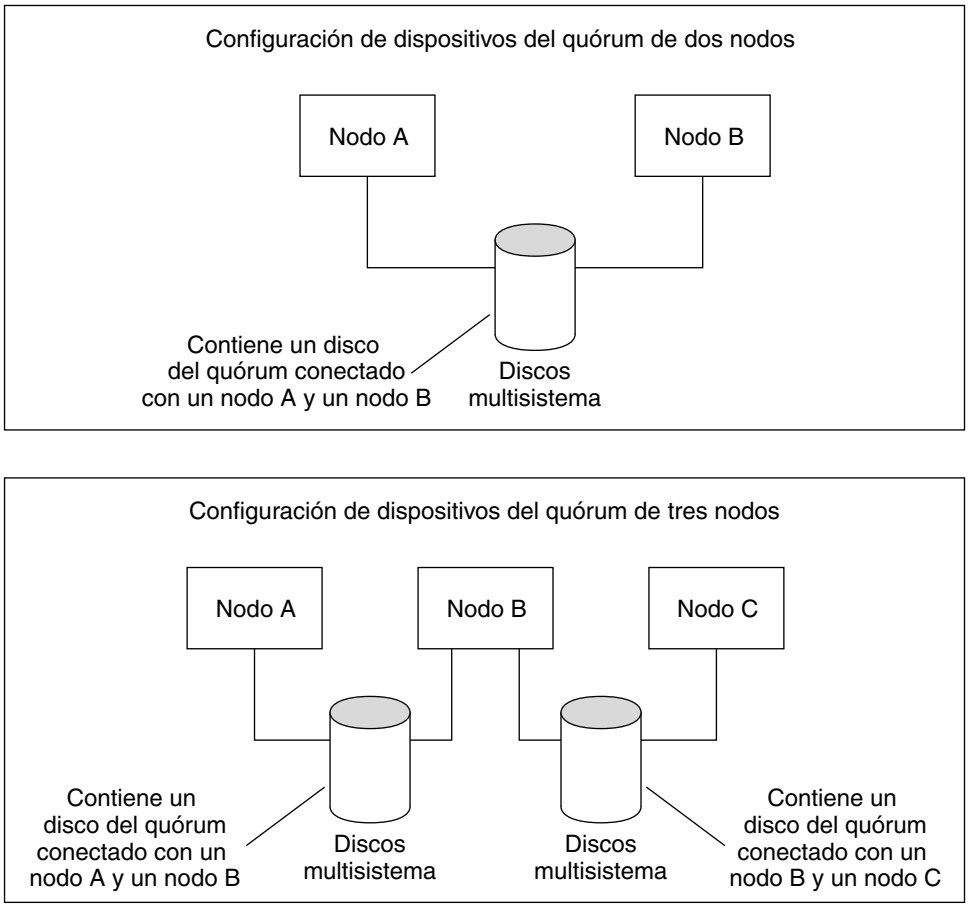


FIGURA 3–3 Ejemplos de configuraciones de los dispositivos del quórum

Directrices del quórum

Use las directrices siguientes cuando configure dispositivos del quórum:

- Establezca un dispositivo del quórum entre todos los nodos que estén conectados al mismo alojamiento de almacenamiento en disco compartido. Agregue un disco dentro del alojamiento compartido como dispositivo del quórum para asegurarse de que si falla cualquier nodo, los otros puedan mantener el quórum y servir a los

grupos de dispositivos de disco del alojamiento compartido.

- El dispositivo del quórum se debe conectar con al menos dos nodos.
- Un dispositivo del quórum puede ser cualquier disco de tipo SCSI-2 o SCSI-3, usado como dispositivo del quórum dual. Los discos conectados a más de dos nodos deben admitir la función Persistent Group Reservation (PGR) aunque no se esté usando el disco como dispositivo del quórum. Para obtener más información, consulte el capítulo sobre planificación en *Sun Cluster 3.1: Guía de instalación del software*.
- Puede usar como dispositivo del quórum un disco que contenga datos del usuario.

Aislamiento de fallos

Un problema fundamental de los clústers es un fallo que provoque en éstos una partición (denominada *esquizofrenia*). Cuando 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. En consecuencia, varios nodos intentando escribir en los discos podrían provocar la corrupción de 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.

Los servicios de dispositivos de disco ofrecen prestaciones para servicios que utilizan discos multisistema. Cuando un miembro del clúster que sirve actualmente como primario (propietario) del grupo de dispositivos de disco cae o deja de ser accesible, se elije otro primario, lo que permite que continúe el acceso al grupo de dispositivos de disco con la mínima interrupción. Durante este proceso, el primario antiguo debe renunciar al acceso a los dispositivos antes de que el nuevo primario pueda iniciarse. Sin embargo, cuando un miembro se descuelga del clúster y deja de estar disponible, éste no puede informar al nodo que libere los dispositivos para los que era primario. Por tanto, se necesita un medio para permitir que los miembros supervivientes tomen control y accedan a los dispositivos globales de los miembros fallidos.

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

Las reservas de los discos SCSI-2 admiten un tipo que concede acceso a todos los nodos conectados al disco (cuando no hay ninguna reserva vigente) o restringe el acceso a un nodo individual (el nodo que retiene la reserva).

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 otro nodo acceda a los discos compartidos. Cuando se produce este aislamiento de fallos es normal que el nodo aislado entre en pánico con mensajes de "conflicto de reserva" en la consola.

El conflicto de reserva se produce porque después de haberse detectado que el nodo ya no es miembro del clúster, se pone una reserva SCSI en todos los discos que están compartidos entre este nodo y los demás nodos. El nodo podría no advertir que se le está aislando y si intenta acceder a uno de los discos compartidos, detecta la reserva y entra en situación de pánico.

Mecanismo de recuperación rápida para aislamiento de fallos

El mecanismo por el que la estructura del clúster se asegura de que un nodo fallido no pueda rearrancar y empezar a escribir en almacenamiento compartido se denomina *recuperación rápida*.

Los nodos que son miembros del clúster habilitan permanentemente un ioctl específico, MHIOCENFAILFAST, para los discos a los que tienen acceso, incluidos los de quórum. ioctl es una directiva para el controlador de disco y da a un nodo la posibilidad de entrar en pánico por sí mismo si éste no puede acceder al disco debido a que otro nodo lo está reservando.

ioctl MHIOCENFAILFAST hace que el controlador compruebe el retorno de error de cada lectura y escritura que emite un nodo al disco cuando busca el código de error `Reservation_Conflict`. ioctl emite periódicamente en segundo plano una operación de prueba al disco para comprobar si se produce `Reservation_Conflict`. Las rutas del flujo de control de segundo y primer plano entran en pánico si se devuelve `Reservation_Conflict`.

En discos SCSI-2, las reservas no son persistentes, pues no resisten los rearranques de los nodos. En los discos SCSI-3 con reserva de grupo persistente (PGR), la información de reserva se almacena en el disco y permanece entre los rearranques de los nodos. El mecanismo de recuperación rápida funciona igual aunque disponga de discos SCSI-2 o SCSI-3.

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 forme parte de la partición que pueda conseguir el quórum coloca reservas en los discos compartidos y cuando el nodo que no tiene el quórum intenta acceder a éstos recibe un conflicto de reserva y entra en condición de pánico como resultado del mecanismo de recuperación rápida.

Después de la condición de pánico el nodo podría rearrancar e intentar volver a unirse al clúster o permanecer en el indicador OpenBoot PROM (OBP). La acción que se lleva a cabo la determina el valor del parámetro `auto-boot?` en la OBP.

Gestores de volúmenes

El sistema SunPlex usa el software de gestión de volúmenes para aumentar la disponibilidad de los datos gracias a las copias especulares y los discos de repuesto en marcha y para manejar fallos y sustituciones de disco.

El sistema SunPlex no tiene su propio componente de gestión de volúmenes interno, sino que funciona de forma coordinada con los siguientes gestores:

- Gestor de volúmenes de Solaris
- VERITAS Volume Manager

El software de gestión de volúmenes del clúster ofrece soporte para:

- Manejo de recuperación de fallos en nodos
- Soporte de ruta múltiple desde distintos nodos
- Acceso remoto transparente a grupos de dispositivos de disco

Cuando los objetos de gestión de volúmenes pasan bajo el control del clúster, se convierten en grupos de dispositivos de disco. Para obtener información sobre los gestores de volúmenes, consulte la documentación que acompañe al software.

Nota – Un requisito importante cuando se planifica los conjuntos o grupos de discos es comprender la asociación de los grupos de dispositivos de disco con los recursos de aplicación (datos) dentro del clúster. Consulte Sun Cluster 3.1: Guía de instalación del software y Sun Cluster 3.1 Data Services Installation and Configuration Guide para ver comentarios sobre estos temas.

Servicios de datos

El término *servicio de datos* describe una aplicación de otras empresas como Oracle o Sun ONE Web Server que se ha configurado para ejecutarse en un clúster en lugar de en un servidor individual. Un servicio de datos se compone de una aplicación, archivos de configuración de Sun Cluster y métodos de gestión Sun Cluster que controlan las acciones siguientes de la aplicación.

- Iniciar
- Detener
- Supervisar y tomar acciones correctivas

En la Figura 3–4 se compara una aplicación que se ejecuta en un servidor de aplicaciones individual (el modelo de servidor individual) con la misma aplicación que se ejecuta en un clúster (el modelo de servidores en clúster). Tenga en cuenta que desde el punto de vista del usuario, no hay diferencia entre las dos configuraciones, excepto en que la aplicación en clúster podría ejecutarse más rápido y con alta disponibilidad.

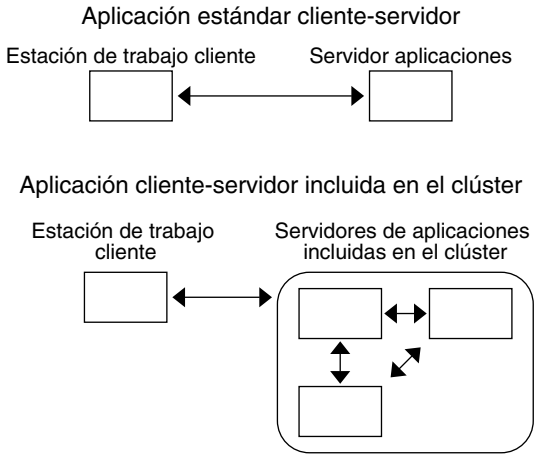


FIGURA 3–4 Configuración cliente/servidor estándar frente a configuración en clúster

En el modelo de servidor individual la aplicación se configura para que acceda al servidor a través de una interfaz de red pública determinada (un nombre de sistema). El nombre de sistema está asociado con este servidor físico.

En el modelo basado en clúster, la interfaz de red pública es un *nombre de sistema lógico* o una *dirección compartida*. El término *recursos de red* se utiliza para los nombres de sistema lógicos y para las direcciones compartidas.

Algunos servicios de datos obligan a especificar nombres de sistema lógicos o direcciones compartidas como interfaces de red, no son intercambiables; otros, en cambio, permiten especificar nombres de sistema lógicos o direcciones compartidas. Consulte la instalación y configuración de cada servicio de datos para obtener los detalles que se deben especificar.

Un recurso de red no está asociado con ningún servidor físico determinado, puede cambiar entre servidores físicos.

Un recurso de red está asociado inicialmente con un nodo, el *primario*. Si éste falla, los recursos de la red y de la aplicación se trasladan a otro nodo del clúster (uno secundario). Cuando el recurso de red se recupera del problema, después de un breve intervalo, el recurso de aplicación continúa ejecutándose en el secundario.

En la Figura 3–5 se compara el modelo de servidor individual con el basado en clúster. Tenga en cuenta que en el modelo de servidor basado en clúster un recurso de red (nombre de sistema lógico, en este ejemplo) puede trasladarse entre dos o más nodos del clúster. La aplicación está configurada para usar este nombre de sistema lógico en lugar de uno asociado a un servidor en particular.

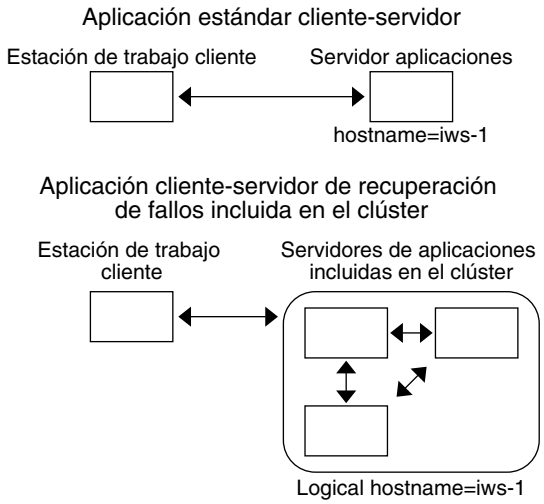


FIGURA 3–5 Comparación entre los nombres de sistema fijo y lógico

Una dirección compartida también está asociada inicialmente con un nodo denominado nodo de interfaz global (GIF). Las direcciones compartidas se usan como interfaces de red únicas con el clúster que se conocen como *interfaces globales*.

La diferencia entre los modelos de nombre de sistema lógico y de servicio escalable es que en éste cada nodo también tiene la dirección compartida configurada activamente en su interfaz de bucle. Esta configuración hace posible tener activas varias instancias de un servicio de datos en varios nodos simultáneamente. El término "servicio escalable" significa que puede agregarse más potencia de CPU a la aplicación agregando nodos del clúster adicionales con lo que el rendimiento se multiplicará.

Si el nodo GIF falla, la dirección compartida puede llevarse a otro nodo que también esté ejecutando una instancia de la aplicación (convirtiendo así al otro nodo en el nuevo GIF). O, la dirección compartida puede trasladarse a otro nodo del clúster que no estuviera ejecutando la aplicación previamente.

En la Figura 3–6 se compara la configuración de servidor individual con la configuración de servicio escalable en clúster. Tenga en cuenta que, en la configuración de servicio escalable, la dirección compartida está presente en todos los nodos. De forma análoga a como se usan los nombres de sistema para los servicios de datos de recuperación de fallos, la aplicación se configura para usar esta dirección compartida en lugar de un nombre de sistema asociado a un servidor determinado.

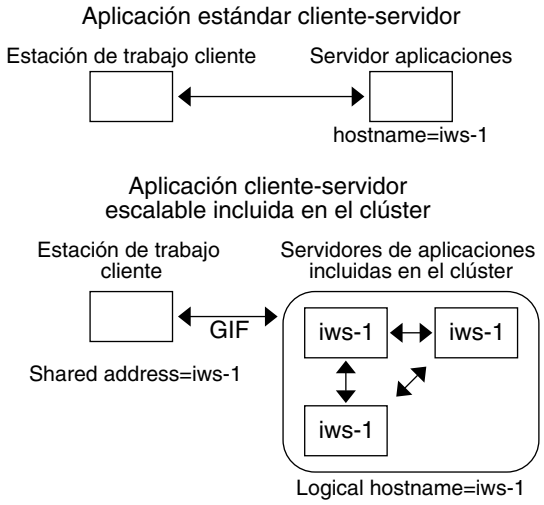


FIGURA 3–6 Comparación entre los nombres de sistema fijos y las direcciones compartidas

Métodos de servicios de datos

El software Sun Cluster proporciona un conjunto de métodos de gestión de servicios que se ejecutan bajo el control del Gestor de grupos de recursos (RGM), el cual los usa para iniciar, detener y supervisar la aplicación en los nodos del clúster. Estos métodos, junto con el software de la estructura del clúster y los discos multisistema, permiten a las aplicaciones convertirse en servicios de datos a prueba de fallos o escalables.

RGM también gestiona los recursos del clúster, incluidas las instancias de una aplicación y de los recursos de red (nombres de sistema lógicos y direcciones compartidas).

Además de los métodos proporcionados por el software Sun Cluster, el sistema SunPlex también proporciona una API y varias herramientas de desarrollo de servicios de datos que permiten a los programadores de aplicaciones desarrollar los métodos de servicio de datos necesarios para que otras aplicaciones se ejecuten como servicios de datos de alta disponibilidad con el software de Sun Cluster.

Servicios de datos a prueba de fallos

Si el nodo en el que se está ejecutando el servicio de datos (el primario) falla, el servicio migra a otro nodo en funcionamiento sin intervención del usuario. Los servicios a prueba de fallos usan un *grupo de recursos de recuperación de fallos*, que es un contenedor para recursos de instancias de aplicaciones y de red (*nombres de sistema 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.

En los servicios de datos a prueba de fallos, las instancias de las aplicaciones sólo se ejecutan en un nodo individual. 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.

Servicios de datos escalables

Los servicios de datos escalables tienen la capacidad de activar instancias en varios nodos; usan dos grupos de recursos: un *grupo de recursos escalable* que contiene los recursos de aplicación y un grupo de recursos a prueba de fallos que contiene los recursos de red (*direcciones compartidas*) de los que depende un 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.

Las peticiones de servicio llegan al clúster a través de una interfaz de red individual (interfaz global) y se distribuyen a los nodos según uno de varios algoritmos definidos por la *política de equilibrio de cargas* que el clúster puede usar para equilibrar la carga del servicio entre varios nodos. Tenga en cuenta que pueden haber varias interfaces globales en distintos nodos alojando otras direcciones compartidas.

En los servicios escalables, las instancias de las aplicaciones se ejecutan en varios nodos simultáneamente. Si el nodo que aloja la interfaz global falla, ésta se traspasa a otro nodo. Si falla una instancia de aplicación en ejecución, ésta intenta reiniciarse en el mismo nodo.

Si no se puede reiniciar una instancia de una aplicación en el mismo nodo, y se configura otro nodo que no esté en uso para ejecutar el servicio, éste se transfiere a este nodo. En caso contrario, continúa ejecutándose en los nodos restantes, lo que podría provocar una degradación en el rendimiento del servicio.

Nota – El estado TCP para cada instancia de aplicación se conserva en el nodo de la instancia, no en el de la interfaz global. Por tanto el fallo en el nodo de interfaz global no afecta a la conexión.

En la Figura 3–7 se muestra un ejemplo de grupo de recursos escalable y las dependencias que existen entre ellos para los servicios escalables. Este ejemplo muestra tres grupos de recursos. El grupo de recursos a prueba de fallos contiene recursos de aplicación que usan los DNS de alta disponibilidad y recursos de red que usan éstos y el servidor Apache Web Server de alta disponibilidad. Los grupos de recursos escalables sólo contienen instancias de la aplicación de Apache Web Server. Tenga en cuenta que existen dependencias de grupos de recursos entre los grupos de

recursos escalables y los a prueba de fallos (líneas sólidas) y que todos los demás recursos de la aplicación de Apache dependen del recurso de red `schost-2`, que es una dirección compartida (líneas punteadas).

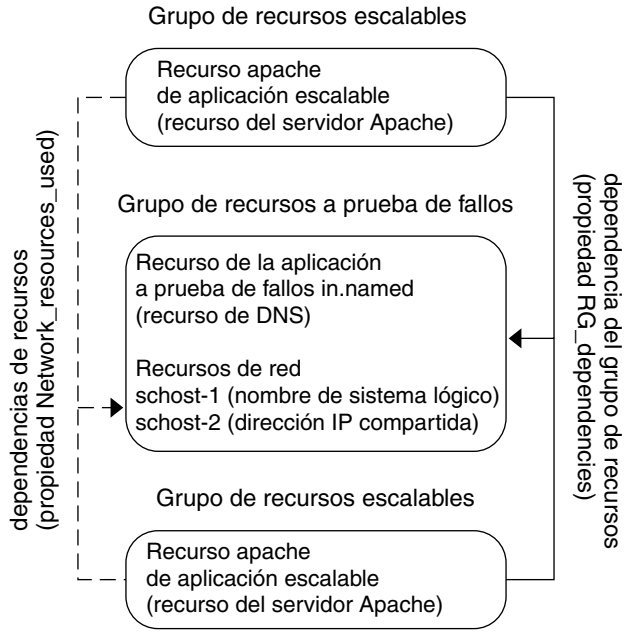


FIGURA 3–7 Ejemplo de los grupos de recursos a prueba de fallos y escalables

Arquitectura de servicio escalable

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. A esto se le denomina servicio de datos escalable. 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. Todas juntas, estas instancias se comportan como si fueran un solo servicio desde el punto de vista el cliente remoto de ese servicio e implementan la funcionalidad del servicio. Por ejemplo, se podría tener un servicio web escalable compuesto de varios daemons `httpd` que se ejecuten en varios nodos. Cualquier daemon `httpd` puede servir una petición de un cliente. El que sirve la solicitud depende de una *política 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.

Un servicio escalable está compuesto por:

- Soporte de infraestructura de red para servicios escalables
- Equilibrio de cargas
- Soporte para servicios de red y datos (usando Gestor de grupos de recursos)

La figura siguiente muestra la arquitectura de servicio escalable.

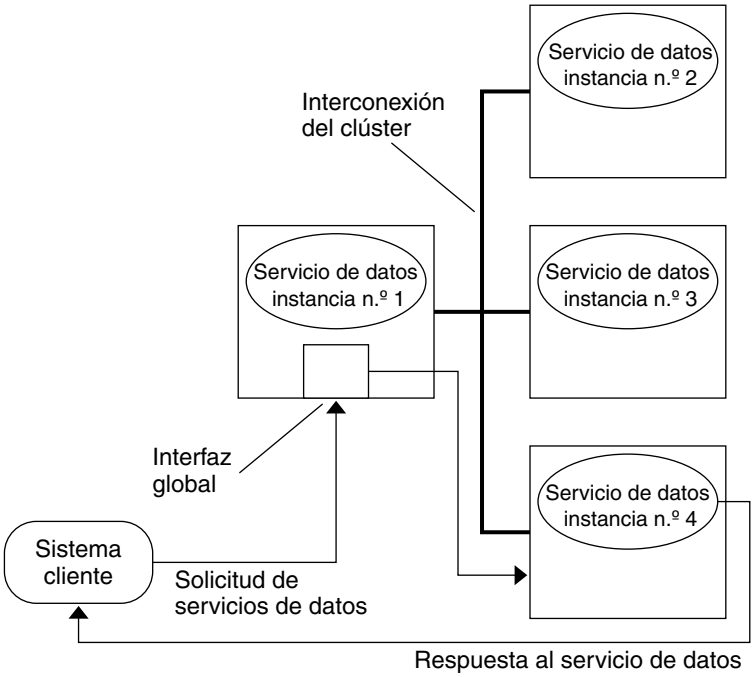


FIGURA 3–8 Arquitectura de servicio escalable

Los nodos que no alojan la interfaz global (nodos delegados) tienen la dirección compartida alojada en sus interfaces de bucle. Los paquetes que van a la interfaz global se distribuyen a otros nodos clúster de acuerdo con las políticas de equilibrio de cargas configurables. Las políticas de equilibrio de cargas posibles se describen a continuación.

Políticas de equilibrio de cargas

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

Existen dos clases de servicios de datos escalables: *puros* y *adosados*. Un servicio puro es aquel cuyas instancias pueden responder a peticiones de clientes sin restricciones. Un servicio adosado es uno en el que un cliente envía peticiones a la misma instancia. 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 supongamos que cada nodo pesa una unidad. Cada nodo dará servicio a 1/3 de las peticiones de cualquier cliente en nombre de ese servicio. El administrador puede cambiar los pesos en todo momento con la interfaz de comandos `scrgadm(1M)` o con la GUI de administración de SunPlex.

Existen dos variedades de servicios adosados, *normales* y *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. Al cliente se le denomina "adosado" con respecto a la instancia del servidor que está a la escucha 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.

Por ejemplo, un explorador web en el cliente se conecta a una dirección IP compartida en el puerto 80 usando tres conexiones TCP distintas, pero las conexiones están intercambiando información de sesión en memoria intermedia entre ellas en el servicio.

Una generalización de política de adosamiento se aplica a varios servicios escalables que intercambien información de forma no visible en la misma instancia. Cuando estos servicios intercambian información de sesión de forma no visible en la misma instancia, se dice que el cliente está "adosado" con respecto a varias instancias de servidor del mismo nodo que está a la escucha en puertos distintos.

Por ejemplo, un cliente de una sede de comercio electrónico rellena su carro de la compra con artículos usando HTTP normal en el puerto 80, pero cambia a SSL en el puerto 443 para enviar datos seguros y pagar por tarjeta de crédito los artículos del carro de la compra.

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.

Un buen ejemplo de esta política es el FTP de modalidad pasiva. Un cliente se conecta a un servidor FTP en el puerto 21 y después el servidor le informa que vuelva a conectarse a un servidor de puertos que está a la escucha en el rango de puertos dinámico. Todas las solicitudes para esta dirección IP se redirigen al mismo nodo que el servidor informó al cliente a través de la información de control.

Tenga en cuenta que además de estas políticas de adosamiento también está vigente de forma predeterminada la política de equilibrio de cargas ponderado, por tanto la petición inicial de un cliente se dirige a la instancia que dicta el repartidor de cargas.

Cuando el cliente ha establecido una afinidad para el nodo en que se está ejecutando la instancia, las futuras peticiones se dirigen a esa instancia siempre que el nodo esté accesible y la política de equilibrio de cargas no cambie.

A continuación se explican detalles adicionales sobre las políticas de equilibrio de cargas específicas.

- Ponderada: la carga se distribuye entre varios nodos según valores de peso especificados. Esta política se establece mediante el valor `LB_WEIGHTED` para la propiedad `Load_balancing_weights`. Si no se establece explícitamente el peso de un nodo, éste toma de forma predeterminada el valor uno.

  La política de ponderación redirije cierta parte del tráfico de los clientes a un nodo determinado. Dado X=peso y A=peso total de todos los nodos activos, cada uno puede esperar recibir aproximadamente X/A del total de conexiones nuevas, cuando éste sea lo bastante grande. Esta política no trata de peticiones individuales.

  Tenga en cuenta que esta política no es de tipo giratoria. Una política de puerta giratoria siempre haría que todas las peticiones de un cliente fueran a un nodo distinto: la primera petición al nodo 1, la segunda al nodo 2, etc.
- Adosada: en esta política el conjunto de puertos es conocido en el momento en el que se configuran los recursos de la aplicación. Se establece mediante el valor `LB_STICKY` para la propiedad del recurso `Load_balancing_policy`.
- Adosada-comodín: esta política tiene menos limitaciones que la "adosada" normal. En un servicio escalable identificado por su dirección IP, el servidor asigna los puertos (y no se conocen de antemano). Los puertos podrían cambiar. Esta política se establece mediante el valor `LB_STICKY_WILD` para la propiedad del recurso `Load_balancing_policy`.

Valores de retroceso

Los grupos de recurso se cubren entre nodos. Cuando esto se produce, el que hacía el papel de secundario se convierte en el nuevo primario. La configuración de retroceso especifica las acciones que acontecen cuando el primario original vuelve a estar en línea. Las opciones son hacer que el primario original se convierta de nuevo en primario (retroceso) o permitir que el que haya tomado su papel continúe con él. La opción deseada se especifica mediante el valor `Failback` de la propiedad del grupo de recurso.

En algunos casos si el nodo original que aloja al grupo de recursos está fallando y rearrancando repetidas veces, configurar el retroceso podría producir una disponibilidad menor para el grupo de recursos.

Supervisores de fallos de servicios de datos

Cada servicio de datos de SunPlex proporciona un supervisor de fallos que explora periódicamente el servicio de datos para determinar su buen estado. Un supervisor de fallos comprueba que los daemons de las aplicaciones estén funcionando y que se esté dando servicio a los clientes. De acuerdo con la información que devuelven las sondas, se pueden tomar acciones predefinidas, como reiniciar daemons o producir una recuperación de fallos.

Desarrollo de nuevos servicios de datos

Sun proporciona archivos de configuración y plantillas de métodos de gestión que permiten hacer funcionar varias aplicaciones como servicios a prueba de fallos o escalables dentro de un clúster. Si Sun no ofrece en algún momento una aplicación que se desee ejecutar como servicio a prueba de fallos o escalable, se puede usar la API DSET u otra para configurar la aplicación a fin de que se ejecute como un servicio de tipo a prueba de fallos o escalable.

Existe un conjunto de criterios para determinar si una aplicación puede convertirse en un servicio a prueba de fallos. Los criterios específicos se describen en los documentos de SunPlex que tratan de las API que se puede usar para las aplicaciones.

Aquí, presentamos algunas directrices que permitirán comprender si el servicio puede aprovechar las ventajas de la arquitectura de servicios de datos escalable. Repase el apartado «Servicios de datos escalables» en la página 61 para obtener más información sobre los servicios escalables.

Los servicios nuevos que sigan estas directrices pueden utilizar las ventajas de los servicios escalables. Si un servicio existente no sigue estas directrices exactamente, será necesario reescribir partes del mismo para que las cumpla.

Un servicio de datos escalable tiene las características siguientes. En primer lugar, un servicio así está compuesto por una o más *instancias* de servidor, cada una de las cuales se ejecuta en un nodo distinto del clúster. Dos o más instancias del mismo servicio no pueden ejecutarse en el mismo nodo.

En segundo lugar, si el servicio ofrece un almacenamiento de datos lógico externo, debe sincronizarse el acceso simultáneo a este almacenamiento desde varias instancias de servidores para evitar perder actualizaciones o leer los datos mientras se estén modificando. Tenga en cuenta que se llama “externo” para distinguir el almacenamiento de estado en memoria y “lógico” porque parece como una entidad única, aunque puede estar replicado. Además, este almacenamiento de datos lógico tiene la propiedad de que cuando alguna instancia de servidor lo actualiza, las demás instancias pueden ver las modificaciones .

El sistema SunPlex proporciona ese almacenamiento externo a través de su sistema de archivos del clúster y las particiones a bajo nivel globales. Como ejemplo, supongamos que un servicio escribe datos nuevos a un archivo de registro cronológico externo o

modifica los datos que haya. Cuando se ejecuten varias instancias de este servicio, cada una tendrá acceso a este registro cronológico externo y podrá intentar acceder al mismo simultáneamente. Cada instancia debe sincronizar su acceso a este registro o de lo contrario se interferirán entre ellas. El servicio podría usar bloqueos de archivo normales de Solaris con `fcntl(2)` y `lockf(3C)` para conseguir la sincronización deseada.

Otro ejemplo de este tipo de almacenamiento es una base de datos de componente trasero como Oracle de alta disponibilidad o Oracle Parallel Server/Real Application Clusters. Tenga en cuenta que este tipo de servidor de base de datos de componente trasero proporciona una sincronización incorporada gracias a las consultas de base de datos o las transacciones de actualización, por lo que no es necesario que las distintas instancias de servidor implementen su propia sincronización.

Un ejemplo de servicio que no es de tipo escalable en su encarnación actual es el servidor IMAP de Sun. El servicio actualiza un almacenamiento, pero es privado y cuando varias instancias de IMAP escriben en él, se sobrescriben porque las actualizaciones no están sincronizadas. El servidor IMAP debe volver a escribirse para que se sincronice el acceso simultáneo.

Finalmente, tenga en cuenta que algunas instancias pueden tener datos privados que sean contradictorios con los de otras. En tal caso, no es necesario que el servicio se preocupe de sincronizar el acceso simultáneo, ya que los datos son privados y sólo esa instancia puede manipularlos. En este caso, se ha de tener cuidado de no almacenar estos datos privados en el sistema de archivos del clúster, porque existe la posibilidad de que resulte accesible globalmente.

API de servicio de datos y de biblioteca de desarrollo de servicio de datos

El sistema SunPlex para que las aplicaciones estén altamente disponibles ofrece:

- Servicios de datos proporcionados como parte del sistema SunPlex
- Una API de servicio de datos
- Una API de biblioteca de desarrollo de servicio de datos
- Un servicio de datos “genérico”

El manual *Sun Cluster 3.1 Data Services Installation and Configuration Guide* describe cómo instalar y configurar los servicios de datos que se proporcionan con el sistema SunPlex. El manual *Sun Cluster 3.1: Guía del desarrollador de los servicios de datos* describe cómo instrumentar otras aplicaciones para hacerlas de alta disponibilidad bajo la estructura de Sun Cluster.

Las API de Sun Cluster permiten a los programadores de aplicaciones desarrollar supervisores de fallos y secuencias que inician y detienen instancias de servicios de datos. Con estas herramientas una aplicación puede instrumentalizarse para que sea

un servicio a prueba de fallos o escalable. Además, el sistema SunPlex ofrece un servicio de datos “genérico” que se puede usar para generar rápidamente métodos de inicio y detención necesarios para la aplicación y ejecutarla como servicio a prueba de fallos o escalable.

Uso de la interconexión del clúster para el tráfico de servicio de datos

Un clúster debe tener varias conexiones de red entre los nodos que formen la interconexión del clúster. El software de agrupación en clúster usa varias interconexiones para la alta disponibilidad y para mejorar el rendimiento. Para el tráfico interno (por ejemplo, datos del sistema de archivos o datos de servicios escalables), los mensajes se dividen entre todas las interconexiones disponibles como si pasaran a través de una puerta giratoria.

La interconexión del clúster también está disponible para aplicaciones, para comunicaciones de alta disponibilidad entre nodos. Por ejemplo, una aplicación distribuida podría tener componentes que se ejecutaran en distintos nodos y que necesitasen comunicarse. Al usar la interconexión del clúster en lugar del transporte público, éstas conexiones pueden soportar el fallo de un enlace individual.

Para usar la interconexión del clúster para la comunicación entre nodos, una aplicación necesita los nombres de sistema privados configurados cuando se instaló el clúster. Por ejemplo, si el nombre de sistema privado para el nodo 1 es `clusternode1-priv`, use ese nombre para comunicarse por la interconexión del clúster con el nodo 1. Los zócalos TCP abiertos mediante este nombre se dirigen por la interconexión del clúster y pueden re-dirigirse de forma transparente en caso de fallo de la red.

Tenga en cuenta que, debido a que los nombres de sistema privados pueden configurarse durante la instalación, la interconexión del clúster podría usar cualquiera de los nombres elegidos en ese momento. El nombre real podrá obtenerse con `scha_cluster_get`(3HA) con el argumento `scha_privatelink_hostname_node`.

Para el uso de la interconexión del clúster a nivel de aplicación, se usa una interconexión simple entre cada par de nodos, aunque si es posible es mejor utilizar interconexiones distintas para distintos pares de nodos. Por ejemplo, consideremos una aplicación que se ejecuta en tres nodos y se comunica mediante la interconexión del clúster. La comunicación entre los nodos 1 y 2 podría darse en la interfaz `hme0`, mientras que la comunicación entre los nodos 1 y 3 podría producirse en la interfaz `qfe1`. Es decir, que la comunicación entre dos nodos estaría limitada a la interconexión simple, mientras que internamente la comunicación del clúster se dividiría entre todas las interconexiones.

Tenga en cuenta que la aplicación comparte la interconexión con el tráfico interno del clúster, por lo que el ancho de banda disponible para la aplicación depende del que se use para el resto del tráfico del clúster. En caso de fallo, el tráfico interno puede desviarse por el resto de interconexiones, mientras que las conexiones de aplicación de una interconexión fallida se cambian a otra que funcione.

Hay dos tipos de direcciones que admiten la interconexión del clúster y `gethostbyname`(3N) en un nombre de sistema privado normalmente devuelve dos direcciones IP. La primera dirección se denomina *dirección pairwise lógica* y la segunda *dirección pernode lógica*.

A cada par de nodos se le asigna una dirección lógica pairwise independiente. Esta pequeña red lógica admite la recuperación de fallos de las conexiones. A cada nodo también se le asigna una dirección pernode fija. Es decir, las direcciones pairwise lógicas para `clusternode1-priv` son distintas para cada nodo, mientras que la dirección pernode lógica para `clusternode1-priv` es la misma en todos los nodos. Un nodo no dispone de dirección pairwise consigo mismo, por tanto `gethostbyname`(`clusternode1-priv`) en el nodo 1 sólo devuelve la dirección pernode lógica.

Tenga en cuenta que las aplicaciones que aceptan conexiones a través de la interconexión del clúster y después verifican la dirección IP por motivos de seguridad deben comprobarse frente a todas las direcciones IP que haya devuelto `gethostbyname` y no sólo con la primera de ellas.

Si necesita direcciones IP uniformes en su aplicación en todos los puntos, configure la aplicación para vincularla con la dirección pernode en los lados del cliente y el servidor para que parezca que todas las conexiones vengan y vayan de la dirección pernode.

Recursos: definición, grupos y tipos

Los servicios de datos usan varios tipos de de *recursos*: las aplicaciones como Apache Web Server o Sun ONE 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.

Los servicios de datos son tipos de recursos. Por ejemplo, Sun Cluster HA para Oracle es el tipo de recurso `SUNW.oracle-server` y Sun Cluster HA for Apache es `SUNW.apache`.

Un recurso es una concreción de *tipo de recurso* que está definida a nivel del clúster. Hay definidos distintos tipos de recursos.

Los recursos de red son tipos de recurso `SUNW.LogicalHostname` o `SUNW.SharedAddress`. Ambos están registrados previamente por el software de Sun Cluster.

Los tipos de recurso `SUNW.HAStorage` y `HAStoragePlus` se usan para sincronizar el inicio de recursos y los grupos de dispositivos de disco de los que éstos dependen. Aseguran que antes de que se inicie un servicio de datos, estén disponibles las rutas de acceso a los puntos de montaje de los sistemas de archivos del clúster, dispositivos globales y grupos de dispositivos. Para obtener más información, consulte "Synchronizing the Startups Between Resource Groups and Disk Device Groups" en el manual *Data Services Installation and Configuration Guide*. (El tipo de recurso `HAStoragePlus` estuvo disponible en Sun Cluster 3.0 5/02 e incorporó otra característica, permitiendo a los sistemas de archivo locales que estuvieran altamente disponibles. Para obtener más información sobre esta función, consulte «Tipo de recurso `HAStoragePlus`» en la página 47.)

Los recursos gestionados por RGM se sitúan en grupos denominados *grupos de recursos*, de forma que pueden ser gestionados como una unidad. El grupo de recursos migra como unidad si se inicia una recuperación de fallos o un cambio.

Nota – Cuando se trae un grupo de recursos que contiene recursos de aplicación en línea, la aplicación se inicia. El método de inicio del servicio de datos espera hasta que la aplicación esté completamente en marcha antes de salir con éxito. Determinar cuándo la aplicación está completamente en marcha sigue el mismo proceso que el supervisor de fallos utiliza para saber que un servicio de datos está ofreciendo servicio a clientes. Consulte el manual *Sun Cluster 3.1 Data Services Installation and Configuration Guide* para obtener más información sobre este proceso.

Gestor de grupos de recursos (RGM)

RGM controla los servicios de datos (aplicaciones) como recursos, que las implementaciones de *tipos de recursos* gestionan. Éstas las proporciona Sun o las crea un desarrollador con una API biblioteca de desarrollo de servicios de datos (API DSDL) o con una API de gestión de recursos (RMAPI). El administrador del clúster crea y gestiona recursos en contenedores llamados *grupos de recursos*. RGM para e inicia los grupos de recursos en nodos seleccionados como respuesta a cambios en la pertenencia al clúster.

RGM actúa en *recursos* y *grupos de recursos*.Las acciones de RGM hacen que los recursos y los grupos de recursos cambien entre los estados en línea y fuera de línea. En el apartado «Estados y configuración de recursos y grupos de recursos » en la página 71 puede encontrarse una descripción completa de los estados y valores que pueden aplicarse a recursos y grupos de recursos. Consulte «Recursos: definición, grupos y tipos» en la página 69 para obtener información sobre cómo ejecutar un proyecto de gestión de recursos bajo el control del RGM.

Estados y configuración de recursos y grupos de recursos

Los administradores aplican a recursos y grupos de recursos configuraciones estáticas que sólo pueden cambiarse con acciones administrativas. RGM cambia los grupos de recursos de "estados" dinámicos. Estos valores y estados se describen en la lista siguiente.

- Gestionados o no gestionados: son valores que afectan a todo el clúster y sólo se aplican a grupos de recursos. Los grupos de recursos los gestiona RGM. El comando `scrgadm(1M)` se puede utilizar para provocar que RGM se encargue o no de la gestión de un grupo de recursos. Estos valores no cambian con la reconfiguración del clúster.

  Cuando se crea un grupo de recursos por primera vez, no se gestiona. Debe gestionarse antes de que cualquier recurso que se incluya en el grupo pueda activarse.

  En algunos servicios de datos, por ejemplo en un servidor web escalable, el trabajo debe hacerse antes de iniciar los recursos de red y después de que se detengan. Este trabajo se hace con los métodos de servicio de datos de inicialización (INIT) y finalización (FINI). Los métodos INIT sólo se ejecutan si el grupo de recursos en el que éstos residen está en estado gestionable.

  Cuando un grupo de recursos cambia de no gestionado a gestionado, los métodos INIT registrados para el grupo se ejecutan en todos los recursos.

  Cuando un grupo de recursos cambia de gestionado a no gestionado, todos los métodos FINI registrados se ejecutan para realizar una limpieza.

  El uso más común de los métodos INIT y FINI son para recursos de red de servicios escalables, pero pueden usarse para cualquier trabajo de inicialización o limpieza que no realice la propia aplicación.
- Habilitados o inhabilitados: son los valores a nivel del clúster que se aplican a los recursos. El comando `scrgadm(1M)` se puede usar para habilitar o inhabilitar los recursos. Estos valores no cambian con la reconfiguración del clúster.

  El valor normal para un recurso es que esté habilitado y funcionando activamente en el sistema.

  Si por algún motivo desea que el recurso no esté disponible en todos los nodos del clúster, puede inhabilitarlo. De esta manera dejará de estar disponible para uso general.
- En línea o fuera de línea: son estados dinámicos y se aplican a recursos y grupos de recursos.

  Estos estados cambian a medida que el clúster pasa a través de los pasos de reconfiguración durante el cambio o la recuperación de fallos. También pueden cambiarse a través de acciones de administración. El comando `scswitch(1M)` se puede usar para cambiar los estados en línea o fuera de línea de un recurso o grupo de recursos.

  Un recurso o un grupo de recursos a prueba de fallos sólo pueden estar en línea en un nodo simultáneamente. Un recurso o un grupo de recursos escalables pueden estar en línea en algunos nodos y fuera de línea en otros. Durante un cambio o una

recuperación de fallos, los grupos de recursos y los recursos que contienen se ponen en fuera de línea en un nodo y después se vuelven a poner en línea en otro distinto.

Si un grupo de recursos está fuera de línea significa que todos sus recursos también lo están. Si un grupo de recursos está en línea significa que todos sus recursos habilitados también lo están.

Los grupos de recursos pueden contener varios recursos, pudiendo haber dependencias entre ellos que requieren que los recursos se pongan en línea y fuera de línea en un orden determinado. Los métodos usados para poner los recursos en línea y fuera de línea pueden ocupar tiempos distintos en cada uno de ellos. Debido a las dependencias de recursos y a las diferencias de tiempo de inicio y finalización, los recursos de un mismo grupo de recursos pueden tener estados de puesta en línea y fuera de línea distintos durante una reconfiguración del clúster.

Propiedades de recursos y grupos de recursos

En los servicios de datos de SunPlex se pueden configurar valores de propiedad para recursos y grupos de recursos. Las propiedades estándar son comunes a todos los servicios de datos. Las propiedades de extensión son específicas de cada servicio de datos. Algunas propiedades estándar y de extensión se configuran con valores predeterminados para que no se tengan que modificar. Otras necesitan configurarse como parte del proceso de creación y configuración de recursos. La documentación de cada servicio de datos especifica qué propiedades de recurso pueden establecerse y cómo hacerlo.

Las propiedades estándar se usan para configurar propiedades de recursos y grupos de recursos que normalmente son independientes de todos los servicios de datos. El conjunto de propiedades estándar se describe en un apéndice de *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Las propiedades de extensión de RGM ofrecen información como la ubicación de los binarios de la aplicación y los archivos de configuración. Las propiedades de extensión se modifican a medida que se configuran los servicios de datos. El conjunto de propiedades de extensión se describe en el capítulo específico para servicios de datos de *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Configuración del proyecto de servicios de datos

Los servicios de datos pueden configurarse para que se ejecuten bajo un nombre de proyecto de Solaris cuando se pongan en línea con RGM. La configuración asocia un recurso o un grupo de recursos gestionados por RGM con un OD de proyecto de Solaris. La correlación desde el recurso o grupo de recursos a un ID de proyecto ofrece la posibilidad de usar controles sofisticados que están disponibles en el entorno Solaris para gestionar las cargas de trabajo y el consumo dentro del clúster.

Nota – Esta configuración sólo se puede realizar si se está ejecutando la versión actual del software de Sun Cluster con Solaris 9.

La funcionalidad de la gestión de Solaris en un entorno clúster permite dar prioridad a las aplicaciones más importantes al compartir un nodo con otras aplicaciones. Las aplicaciones podrían compartir un nodo si hay servicios consolidados o porque se haya producido una recuperación de fallos. El uso de la funcionalidad de la gestión que se describe aquí podría mejorar la disponibilidad de una aplicación crítica ya que evita que otras aplicaciones de prioridad baja consuman demasiados suministros del sistema, como tiempo de CPU.

Nota – La documentación de Solaris sobre esta función describe el tiempo de CPU, los procesos, las tareas y componentes similares como 'recursos'. Sin embargo, la documentación de Sun Cluster usa el término 'recursos' para describir entidades que están bajo el control del RGM. El apartado siguiente usará el término 'recurso' para hacer referencia a las entidades de Sun Cluster bajo el control del RGM y el término 'suministros' para hacer referencia a tiempo de CPU, procesos y tareas.

Este apartado ofrece una descripción conceptual de la configuración de los servicios de datos para ejecutar procesos en un `project`(4) especificado de Solaris 9. Este apartado también describe varios escenarios de recuperación de fallos y sugerencias para planificar el uso de la funcionalidad de la gestión que ofrece el entorno Solaris. Para obtener documentación detallada sobre conceptos y procedimientos de la función de la gestión, consulte *System Administration Guide: Resource Management and Network Services* en *Solaris 9 System Administrator Collection*.

Al configurar recursos y grupos de recursos para usar la funcionalidad de la gestión de Solaris en un clúster, considere el uso del siguiente proceso de alto nivel:

1. Configure aplicaciones como parte del recurso.
2. Configure recursos como parte de un grupo de recursos.
3. Habilite los recursos del grupo.
4. Haga gestionable el grupo de recursos.
5. Cree un proyecto de Solaris para el grupo de recursos.
6. Configure propiedades estándar para asociar el nombre del grupo de recursos con el proyecto que se ha creado en el paso 5.
7. Ponga en línea el grupo de recursos.

Para configurar las propiedades estándar `Resource_project_name` o `RG_project_name` para asociar el ID de proyecto de Solaris con el recurso o grupo de recursos, use la opción `-y` con el comando `scrgadm`(1M). Configure los valores de

la propiedad al recurso o grupo de recursos. Consulte “Standard Properties” in *Sun Cluster 3.1 Data Services Installation and Configuration Guide*para obtener definiciones de las propiedades. Consulte `r_properties`(5) y `rg_properties`(5) para obtener las descripciones adecuadas.

El nombre del proyecto especificado debe existir en la base de datos de proyectos (`/etc/project`) y el superusuario debe estar configurado como miembro del proyecto con ese nombre. Consulte “Projects and Tasks” in *System Administration Guide: Resource Management and Network Services* en *Solaris 9 System Administrator Collection* para obtener información conceptual sobre el nombre del proyecto. Consulte `project`(4) para obtener una descripción de la sintaxis de los archivos del proyecto.

Cuando RGM pone en línea el recurso o grupo de recursos, ejecuta los procesos relacionados que hay bajo el nombre del proyecto.

Nota – Los usuarios pueden asociar recursos o grupos de recursos con proyectos en cualquier momento. Sin embargo, el nombre del proyecto nuevo no tiene vigencia hasta que el recurso o grupo de recursos se pone fuera de línea y se vuelve a poner en línea usando RGM.

Ejecutar recursos y grupos de recursos bajo el nombre del proyecto permite configurar las funciones siguientes para gestionar suministros de sistema en todo el clúster.

- Contabilidad ampliada: ofrece una forma flexible de registrar el consumo según las tareas o los procesos. La contabilidad ampliada permite examinar el uso histórico y evaluar las necesidades de capacidad para cargas de trabajo futuras.
- Controles: proporcionan un mecanismo para ajustarse a los suministros del sistema. Puede evitarse que los procesos, tareas y proyectos consuman grandes cantidades de suministros de sistema especificados.
- Programación de partición justa (FSS): ofrece la posibilidad de controlar la ubicación de tiempo de CPU disponible entre cargas de trabajo según su importancia. Ésta se mide por el número de particiones de tiempo de CPU que se asigna a cada carga de trabajo. Consulte `dispadmin`(1M) para obtener una descripción de las líneas de comandos para establecer FSS como programador predeterminado. Consulte también `priocntl`(1), `ps`(1) y `FSS`(7) para obtener más información.
- Agrupaciones: ofrecen la posibilidad de usar particiones para aplicaciones interactivas según los requisitos de éstas. La agrupaciones pueden usarse para particionar un servidor que admite distintas aplicaciones de software. El uso de agrupaciones produce una respuesta más predecible para cada aplicación.

Determinación de requisitos para la configuración de proyectos

Antes de que se configuren servicios de datos para usar los controles que ofrece Solaris en un entorno Sun Cluster, es necesario decidir cómo se desean controlar y seguir los recursos en caso de cambios o recuperación de fallos. Considere identificar dependencias dentro del clúster antes de configurar un proyecto nuevo. Por ejemplo, los recursos y grupos de recursos dependen de grupos de dispositivos de disco. Las propiedades del grupo de recursos `nodelist`, `failback`, `maximum_primaries` y `desired_primaries` se configuran con `scrgadm`(1M) para identificar prioridades de las listas de nodos en el grupo de recursos. Consulte "Relationship Between Resource Groups and Disk Device Groups" in *Sun Cluster 3.1 Data Services Installation and Configuration Guide* para obtener una explicación breve sobre las dependencias de la lista de nodos entre los grupos de recursos y los grupos de dispositivos de disco. Para obtener descripciones de propiedad detalladas, consulte `rg_properties`(5).

Utilice las propiedades `preferenced` y `failback` configuradas con `scrgadm`(1M) y `scsetup`(1M) para determinar las prioridades de la lista de nodos del grupo de dispositivos de disco. Para obtener información sobre procedimientos, consulte "Cómo cambiar las propiedades de un dispositivo de disco" en "Administering Disk Device Groups" in *Sun Cluster 3.1 System Administration Guide*. Consulte «Los componentes de hardware del sistema SunPlex» en la página 19 para obtener información conceptual sobre la configuración de los nodos y el comportamiento de los servicios de datos a prueba de fallos y escalables.

Si se configura todos los nodos del clúster de forma idéntica, los límites de uso se hacen cumplir de forma idéntica en los nodos primarios y secundarios. Es necesario que los parámetros de configuración de los proyectos no sean idénticos en todas las aplicaciones de todos los nodos. Todos los proyectos asociados con la aplicación deben ser accesibles al menos por la base de datos de proyectos en todos los maestros potenciales de esa aplicación. Supongamos que la aplicación 1 está controlada por *phys-schost-1* pero podría efectuarse un cambio o resolución de error a *phys-schost-2* o *phys-schost-3*. El proyecto asociado con la aplicación 1 debe estar accesible en los tres nodos (*phys-schost-1*, *phys-schost-2* y *phys-schost-3*).

Nota – La información de la base de datos de proyectos puede ser local `/etc/project` o almacenarse en el mapa NIS o el directorio LDAP.

El entorno Solaris permite una configuración flexible de parámetros de uso y Sun Cluster pone pocas restricciones. Las opciones de configuración dependen de las necesidades de la sede. Considere las directrices generales de los apartados siguientes antes de configurar los sistemas.

Establecimiento de límites de memoria virtual por proceso

Configure el control `process.max-address-space` para limitar la memoria virtual a cada proceso. Consulte `rctladm`(1M) para obtener información detallada sobre la configuración del valor `process.max-address-space`.

Al utilizar controles de gestión con Sun Cluster, se deben configurar los límites de memoria apropiadamente para evitar una recuperación de fallos innecesaria de aplicaciones y un efecto “ping-pong”. En general:

- No configure unos límites de memoria demasiado bajos.

  Cuando una aplicación llegue a su límite de memoria, podría producirse una recuperación de fallos. Esta directriz es especialmente importante para aplicaciones de base de datos, ya que al alcanzar un límite de memoria virtual se pueden producir consecuencias inesperadas.
- No configure límites de memoria idénticos en nodos primarios y secundarios.

  Límites idénticos podrían causar un efecto ping-pong cuando una aplicación alcance su límite de memoria y se recupere el error en un nodo secundario con un límite de memoria idéntico. Configure el límite de memoria un poco por encima en el nodo secundario. La diferencia en los límites de memoria ayuda a evitar la situación ping-pong y da al administrador del sistema un periodo de tiempo en el que ajustar los parámetros como sea necesario.
- Use los límites de memoria de la gestión de recursos para el equilibrio de cargas.

  Por ejemplo, puede usar límites de memoria para evitar que que una aplicación que no funcione bien consuma demasiado espacio de intercambio en disco.

Escenarios de recuperación de fallos

Puede configurar parámetros de gestión para que la asignación en la configuración del proyecto (`/etc/project`) funcione normalmente en el clúster y en las situaciones de cambio o recuperación de fallos.

Los apartados siguientes son escenarios de ejemplo.

- Los dos primeros apartados, “Clúster de dos nodos con dos aplicaciones“ y “Clúster de dos nodos con tres aplicaciones“, muestran escenarios de recuperación de fallos para nodos completos.
- El apartado “Recuperación de fallos sólo de grupos de recursos“ ilustra el funcionamiento de la recuperación de fallos sólo para una aplicación.

En un entorno del clúster las aplicaciones se configuran como parte de un recurso y éste como parte de un grupo de recursos (RG). Cuando se produce un fallo, el grupo de recursos, junto con su aplicación asociada, se transfieren a otro nodo. En los ejemplos siguientes los recursos no se muestran específicamente. Se presupone que cada recurso sólo tiene una aplicación.

Nota – La recuperación de fallos se produce en el orden de lista de nodos especificado en la RGM.

Los ejemplos siguientes tienen estas limitaciones:

- La aplicación 1 (App-1) está configurada en el grupo de recursos RG-1.
- La aplicación 2 (App-2) está configurada en el grupo de recursos RG-2.
- La aplicación 3 (App-3) está configurada en el grupo de recursos RG-3.

Aunque los números de particiones asignadas siguen siendo los mismos, el porcentaje de tiempo de CPU asignado a cada aplicación cambia después de la recuperación de fallos. Este porcentaje depende del número de aplicaciones que se están ejecutando en el nodo y del número de particiones que se han asignado a cada aplicación activa.

En estos escenarios, se asumen las configuraciones siguientes.

- Todas las aplicaciones están configuradas bajo un proyecto común.
- Cada recurso dispone de una única aplicación.
- Las aplicaciones son los únicos procesos activos de los nodos.
- Las bases de datos de los proyectos están configuradas igual en todos los nodos del clúster.

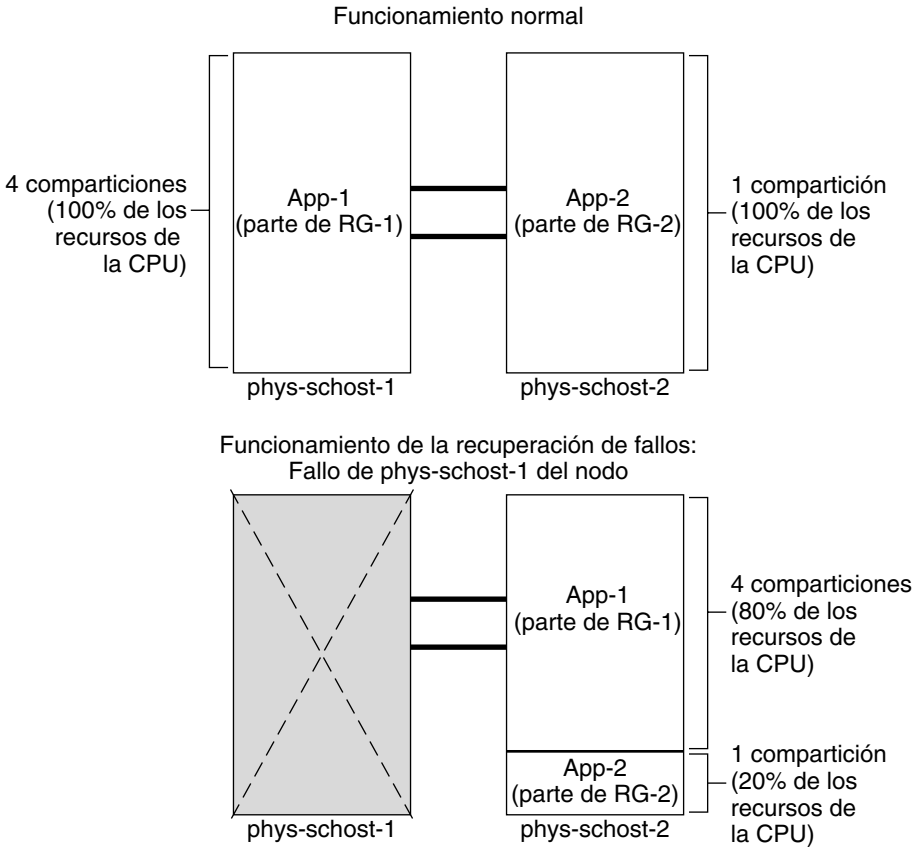
Clúster de dos nodos con dos aplicaciones

Se pueden configurar dos aplicaciones en un clúster de dos nodos para asegurarse de que todos los sistemas físicos (*phys-schost-1*, *phys-schost-2*) actúen como maestros predeterminados de una aplicación. Cada sistema físico actúa como el nodo secundario del otro. Todos los proyectos asociados con las aplicaciones 1 y 2 deben estar representado en los archivos de la base de datos de proyectos de ambos nodos. Cuando el clúster se está ejecutando normalmente, todas las aplicaciones se ejecutan en su principal predeterminado, donde la función de gestión le asigna todo el tiempo de CPU.

Después de una recuperación de fallos o cambio, las dos aplicaciones se ejecutan en un único nodo en que se les asigna particiones, como se especifica en el archivo de configuración. Por ejemplo, esta entrada del archivo `/etc/project` especifica que a la aplicación 1 se le asignan 4 particiones y a la aplicación 2 se le asigna 1 partición.

```
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)
```

El diagrama siguiente ilustra las operaciones normales y de recuperación de fallos de esta configuración. El número de particiones que se asignen no cambia. Sin embargo, el porcentaje de tiempo de CPU disponible para cada aplicación puede cambiar, dependiendo del número de particiones asignadas a cada proceso que requiera tiempo de CPU.



Clúster de dos nodos con tres aplicaciones

En un clúster de dos nodos con tres aplicaciones, se puede configurar un sistema físico (*phys-schost-1*) como principal predeterminado de una aplicación y el segundo sistema físico (*phys-schost-2*) como maestro predeterminado para las otras dos aplicaciones. Consideremos el siguiente archivo de base de datos de proyectos de ejemplo que está en todos los nodos. El archivo de base de datos de proyectos no cambia cuando se produce una recuperación de fallos o un cambio.

```
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)
```

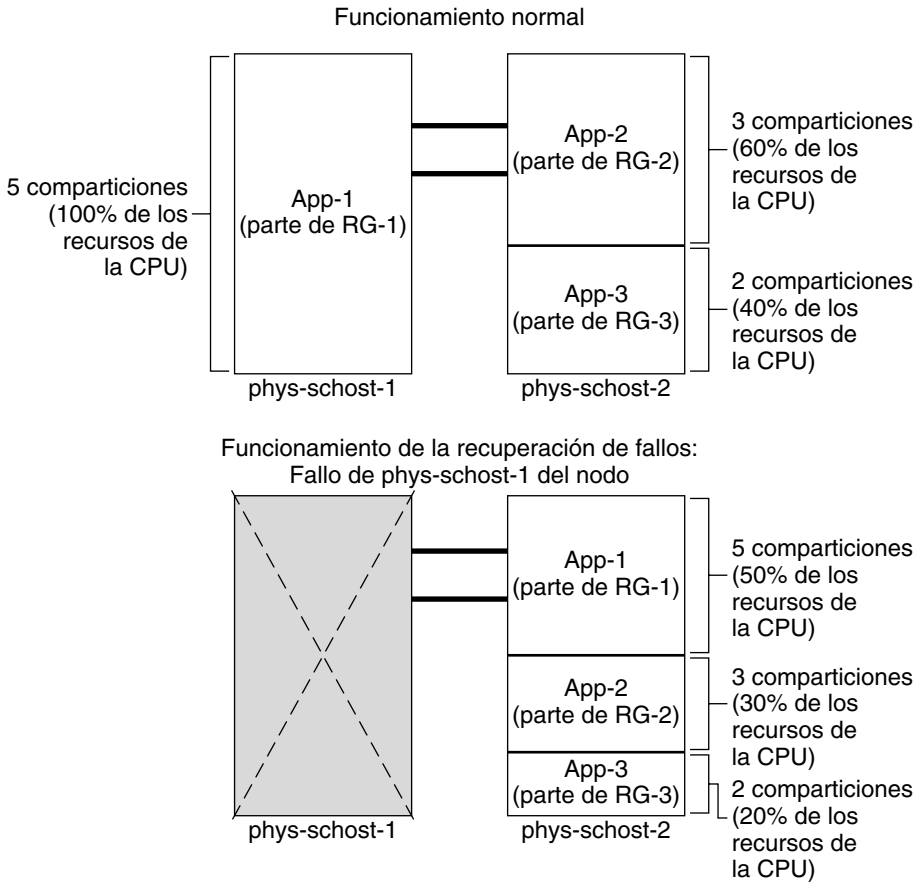
Cuando el clúster se está ejecutando normalmente, a la aplicación 1 se le asignan 5 particiones en su maestro predeterminado, *phys-schost-1*. Este número es equivalente al 100 por cien del tiempo de CPU porque es la única aplicación que lo demanda en ese

nodo. A las aplicaciones 2 y 3 se les asigna 3 y 2 particiones respectivamente en su maestro predeterminado, *phys-schost-2*. La aplicación 2 recibiría el 60 por ciento del tiempo de CPU y la aplicación 3 recibiría el 40 por ciento durante el funcionamiento normal.

Si se produce una recuperación de fallos o un cambio y la aplicación 1 se cambia a *phys-schost-2*, las particiones para las tres aplicaciones siguen siendo las mismas. Sin embargo, los porcentajes de recursos de CPU se reasignan según el archivo de base de datos de proyectos.

- La aplicación 1, con 5 particiones, recibe el 50 por ciento de CPU.
- La aplicación 2, con tres particiones, recibe el 30 por ciento de CPU.
- La aplicación 3, con 2 particiones, recibe el 20 por ciento de CPU.

El diagrama siguiente ilustra el funcionamiento normal y las operaciones de recuperación de fallos de esta configuración.



Recuperación de fallos sólo de grupos de recursos

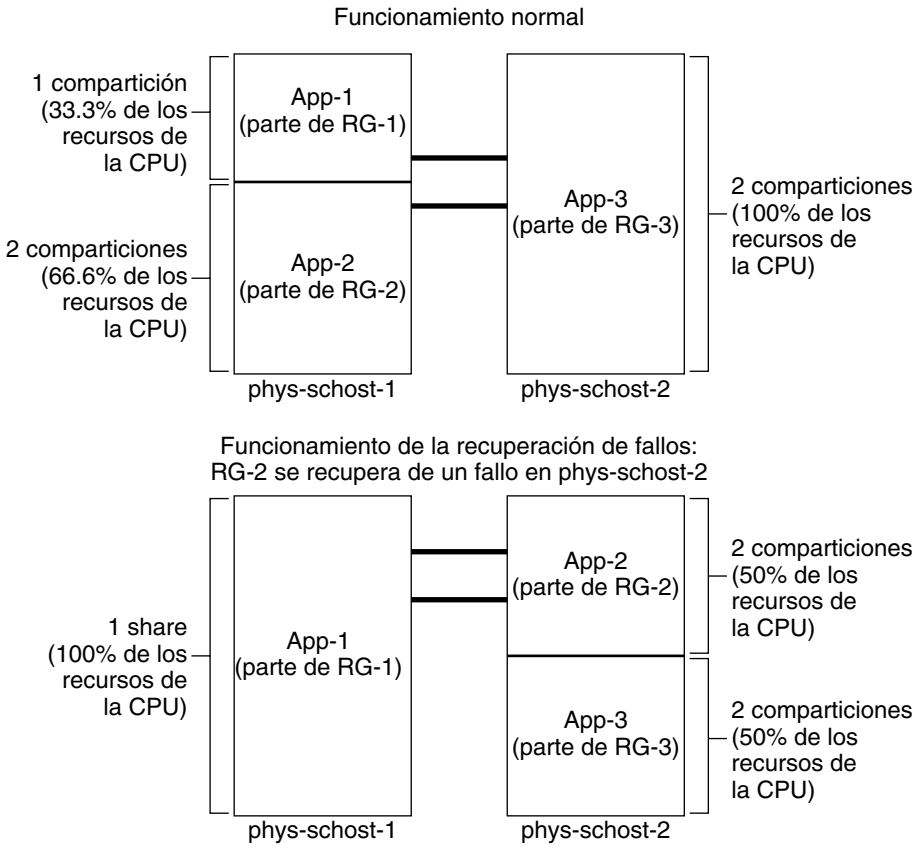
En una configuración en la que varios grupos de recursos tienen el mismo maestro predeterminado, un grupo de recursos (y sus aplicaciones asociadas) pueden entrar en recuperación de fallos o cambiarse a un nodo secundario. Mientras tanto el maestro predeterminado está ejecutándose en el clúster.

Nota – Durante la recuperación de fallos a la aplicación que falla se le asignan recursos, como los que se especifican en el archivo de configuración en el nodo secundario. En este ejemplo, los archivos de la base de datos de proyectos de los nodos primario y secundario tienen las mismas configuraciones.

Por ejemplo, este archivo de configuración de ejemplo especifica que a la aplicación 1 se le asigna 1 partición, a la aplicación 2 se le asignan 2 y a la aplicación 3 se le asignan otras 2.

```
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)
```

El diagrama siguiente ilustra las operaciones normal y de resolución de problemas de esta configuración, en que RG-2, que contiene la aplicación 2, falla sobre *phys-schost-2*. Tenga en cuenta que el número de particiones asignadas no cambia. Sin embargo, el porcentaje de tiempo de CPU disponible para cada aplicación puede cambiar, dependiendo del número de particiones asignadas a cada aplicación que requiera tiempo de CPU.



Adaptadores de red pública y 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 un par de adaptadores.

El software de Solaris Internet Protocol (IP) Network Multipathing en Sun Cluster ofrece el mecanismo básico para supervisar los adaptadores de red pública y cambiar direcciones IP de un adaptador a otro cuando se detecta un fallo. Todos los nodos del clúster tienen su propia configuración Ruta múltiple de red IP, que puede ser distinta de la de otros nodos del clúster.

Los adaptadores de red pública están organizados en *grupos de ruta múltiple IP* (grupos de ruta múltiple). Cada grupo de ruta múltiple tiene 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. El daemon de ruta múltiple `in.mpathd` usa una dirección IP de prueba para detectar fallos y reparaciones, si detecta un fallo en uno de los adaptadores, se produce una recuperación de fallos. Todos los accesos de red se transfieren desde el adaptador erróneo a otro que funcione del grupo de ruta múltiple, con los cual se mantiene la conectividad de red pública para el nodo. Si se configuró una interfaz en espera, el daemon lo eligirá. En caso contrario, `in.mpathd` elige la interfaz con la dirección IP de número inferior. Debido a que la recuperación del fallos se produce en la interfaz del adaptador, las conexiones de mayor nivel como TCP no se resultan afectadas excepto por un breve retraso durante la recuperación de fallos. Cuando se completa satisfactoriamente la recuperación de fallos de direcciones IP, se envían emisiones ARP generalizadas. Así se mantiene la conectividad con clientes remotos.

Nota – Debido a la característica de recuperación de congestiones de TCP, los puntos finales de TCP pueden verse más retrasados después de una recuperación de fallos satisfactoria, ya que algunos segmentos podrían perderse durante el proceso, activándose el mecanismo de control de la congestión en TCP.

Los grupos de ruta múltiple ofrecen bloques de construcción para nombres de sistema lógicos y recursos de dirección compartida. También se pueden crear grupos de ruta múltiple independientemente de los nombres de sistema lógicos y los recursos de dirección compartida para supervisar la conectividad de red pública de los nodos del clúster. El mismo grupo de ruta múltiple de un nodo puede alojar cualquier número de nombres de sistema lógicos o recursos de dirección compartida. Para obtener más información sobre nombres de sistema lógicos y recursos de dirección compartida, consulte *Sun Cluster 3.1 Data Services Installation and Configuration Guide*.

Nota – El diseño del mecanismo Ruta múltiple de red IP está pensado para detectar y filtrar fallos de adaptadores, no para recuperarse de un administrador usando `ifconfig(1M)` para quitar una de las direcciones IP lógicas (o compartidas). El software de Sun Cluster trata las direcciones IP lógicas y compartidas como recursos gestionados por RGM. La forma correcta de que un administrador agregue o retire una dirección IP es usar `scrgadm`(1M) para modificar el grupo de recursos que lo contiene.

Para obtener más información sobre la implementación de Solaris de la Ruta múltiple de red IP, consulte la documentación apropiada para el sistema operativo Solaris instalado en su clúster.

Versión del sistema operativo	Para obtener instrucciones, vaya a
Sistema operativo Solaris 8	*IP Network Multipathing Administration Guide*
Sistema operativo Solaris 9	"IP Network Multipathing Topics" en *System Administration Guide: IP Services*

Soporte de reconfiguración dinámica

El soporte de Sun Cluster 3.1 para la función de software de reconfiguración dinámica (DR) se está desarrollando en fases incrementales. Este apartado describe los conceptos y consideraciones para el soporte de Sun Cluster 3.1 de la función DR.

Tenga en cuenta que todos los requisitos, procedimientos y restricciones que se documentan para la función DR de Solaris también se aplican al soporte DR de Sun Cluster (excepto para el funcionamiento silencioso del sistema operativo). Por consiguiente, se ha de repasar la documentación de la función de DR de Solaris *antes* de utilizarla con el software Sun Cluster. Deberá revisar específicamente las cuestiones que afectan a los dispositivos de E/S que no son de la red durante una operación de desconexión de DR. Están disponibles para su descarga *Sun Enterprise 10000 Dynamic Reconfiguration User Guide* y *Sun Enterprise 10000 Dynamic Reconfiguration Reference Manual* (de las colecciones *Solaris 8 on Sun Hardware* o *Solaris 9 on Sun Hardware*) desde http://docs.sun.com.

Descripción general de la reconfiguración dinámica

La función DR permite operaciones, como la extracción de hardware en sistemas que están en funcionamiento. Los procesos DR están diseñados para asegurar el funcionamiento continuo del sistema sin necesidad de pararlo o interrumpir la disponibilidad del clúster.

DR funciona a nivel de placa, por lo tanto, afecta a todos los componentes de ésta. Cada placa puede contener varios componentes, como CPU, memorias e interfaces periféricas para unidades de disco, unidades de cinta y conexiones en red.

Extraer una placa que contenga componentes activos puede provocar errores en el sistema. Antes de extraer una placa, el subsistema DR consulta otros subsistemas, como Sun Cluster, para determinar si los componentes de la placa se están utilizando. Si el subsistema DR encuentra que la placa se está usando, la operación de extraer la placa por DR no se lleva a cabo. Por tanto, siempre es seguro usar una operación de extracción de placa por DR ya que el subsistema DR rechaza las operaciones en placas que contengan componentes activos.

La operación de agregar placas DR también es siempre segura. El sistema pone en funcionamiento automáticamente las CPU y la memoria de una placa recién añadida. Sin embargo, el administrador del sistema debe configurar manualmente el clúster para usar activamente componentes que están en la placa recién añadida.

Nota – El subsistema DR tiene varios niveles. Si un nivel inferior informa de un error, el superior también informa del mismo error. Sin embargo, cuando el nivel inferior informa del error específico, el superior informará de "Error desconocido". Los administradores del sistema deberían ignorar el "Error desconocido" del que informa el nivel superior.

Los apartados siguientes describen consideraciones de DR para los distintos tipos de dispositivos.

Consideraciones de la agrupación DR para dispositivos de CPU

El software Sun Cluster no rechazará una operación de extracción de placa con DR debido a la presencia de dispositivos de CPU.

Cuando una operación de añadir placa con DR tenga éxito, los dispositivos de CPU de la placa añadida se incorporarán automáticamente al funcionamiento del sistema.

Consideraciones de la agrupación DR para memoria

Con respecto a DR, existen dos tipos de memoria que considerar que difieren sólo en su uso. El hardware real es el mismo para ambos tipos.

La memoria usada por el sistema operativo se llama caja de la memoria del núcleo. El software Sun Cluster no admite operaciones de extracción en placas que contengan la caja de la memoria del núcleo y rechazará cualquiera de estas operaciones. Cuando una operación de extracción de placa con DR pertenezca a una memoria distinta de la caja de la memoria del núcleo, Sun Cluster no rechazará la operación.

Cuando una operación de añadir placa con DR que pertenezca a memoria tenga éxito, la memoria de la placa añadida se incorporará automáticamente al funcionamiento del sistema.

Consideraciones de la agrupación DR para unidades de disco y cinta

Sun Cluster rechaza las operaciones DR de extraer-placa en las unidades activas del nodo principal. Las operaciones de extracción de placa con DR pueden llevarse a cabo en unidades no activas del nodo primario y en cualquier unidad del nodo secundario. Después de la operación de DR, el acceso a los datos del clúster prosigue de la misma forma.

Nota – Sun Cluster rechaza las operaciones DR que afecten a la disponibilidad de los dispositivos del quórum. Para consideraciones sobre los dispositivos del quórum y el procedimiento para llevar a cabo operaciones DR en ellos, consulte «Consideraciones de la agrupación DR para los dispositivos del quórum» en la página 85.

Consulte *Sun Cluster 3.1: Guía de administración del sistema* para obtener instrucciones detalladas sobre cómo llevar a cabo estas acciones.

Consideraciones de la agrupación DR para los dispositivos del quórum

Si la operación de extracción de placa con DR pertenece a una placa que contenga una interfaz con un dispositivo configurado para el quórum, Sun Cluster rechazará la operación e identificará el dispositivo del quórum que podría resultar afectado por ésta. Es necesario inhabilitar el dispositivo como del quórum antes de poder efectuar una operación de extracción de placa con DR.

Consulte *Sun Cluster 3.1: Guía de administración del sistema* para obtener instrucciones detalladas sobre cómo llevar a cabo estas acciones.

Consideraciones de la agrupación DR para interfaces de interconexión del clúster

Si la operación de extracción de placa con DR pertenece a una placa que contenga una interfaz de interconexión del clúster activa, Sun Cluster rechazará la operación e identificará la interfaz que podría resultar afectada por la operación. Es necesario usar una herramienta de administración de Sun Cluster para inhabilitar la interfaz activa antes de que la operación de DR pueda tener éxito (consulte también la nota de precaución siguiente).

Consulte *Sun Cluster 3.1: Guía de administración del sistema* para obtener instrucciones detalladas sobre cómo llevar a cabo estas acciones.

Precaución – Sun Cluster requiere que todos los nodos del clúster tengan al menos una ruta de acceso que funcione con cada uno de los demás nodos del clúster. No inhabilite una interfaz de interconexión privada en el caso de que represente la ultima ruta a cualquiera de los nodos del clúster.

Consideraciones de la agrupación DR para interfaces de red pública

Si la operación de extracción de placa con DR pertenece a una placa que contenga una interfaz de red pública activa, Sun Cluster rechazará la operación e identificará la interfaz que podría resultar afectada por la operación. Antes de extraer una placa que contenga una interfaz de red activa, todo el tráfico de esa interfaz debe cambiarse a otra interfaz funcional del grupo de ruta múltiple mediante el comando `if_mpadm`(1M).

Precaución – Si el resto de adaptadores de red fallan mientras se está efectuando la operación de extracción con DR en el adaptador de red inhabilitado, la disponibilidad se verá afectada. El adaptador restante no tiene a quién transferir el control durante la operación de DR.

Consulte *Sun Cluster 3.1: Guía de administración del sistema* para obtener instrucciones detalladas sobre cómo llevar a cabo operaciones de DR en interfaces de red públicas.

CAPÍTULO 4

Preguntas más frecuentes (FAQ)

Este capítulo incluye las respuestas a las preguntas más frecuentes sobre el sistema SunPlex. Las preguntas están organizadas por temas.

FAQ sobre alta disponibilidad

- **¿Qué es exactamente un sistema de alta disponibilidad?**

  El sistema SunPlex define la alta disponibilidad (HA) como la posibilidad del clúster de mantener en marcha una aplicación aunque se produzca un fallo que en condiciones normales provocaría que el servidor no estuviera disponible.

- **¿Cuál es el proceso por el que el clúster proporciona alta disponibilidad?**

  A través de un proceso conocido como recuperación de fallos, el marco del clúster proporciona un entorno de alta disponibilidad. Recuperación de fallos es una serie de pasos que realiza el clúster para migrar recursos de servicios de datos de un nodo fallido a otro operativo dentro del clúster.

- **¿Cuál es la diferencia entre un servicio de datos a prueba de fallos y otro escalable?**

  Existen dos tipos de servicios de datos de alta disponibilidad, a prueba de fallos y escalable.

  Un servicio de datos a prueba de fallos ejecuta una aplicación sólo en un nodo primario del clúster cada vez. Los demás nodos pueden ejecutar otras aplicaciones, pero cada una se ejecuta sólo en un nodo. Si un nodo primario falla, las aplicaciones que se ejecutan en este nodo se trasladan a otro nodo y continúan ejecutándose.

  Un servicio escalable reparte una aplicación entre varios nodos para crear un servicio lógico único que aprovecha el número de nodos y procesadores de todo el clúster en el que se ejecutan.

Para cada aplicación un nodo aloja la interfaz física para el clúster. Este nodo se denomina interfaz global (GIF). Pueden haber varios nodos GIF en el clúster. Cada uno de ellos aloja una o más interfaces lógicas que pueden usar los servicios escalables y que reciben el nombre de *interfaces globales*. Un nodo GIF aloja una interfaz global para todas las solicitudes hacia una aplicación en particular y las despacha a los distintos nodos en los que se esté ejecutando el servidor de la aplicación. Si el nodo GIF falla, la interfaz global se traslada a un nodo superviviente.

Si falla alguno de los nodos en los que se está ejecutando la aplicación, ésta continúa ejecutándose en los demás nodos con alguna pérdida de rendimiento hasta que el nodo fallido vuelve al clúster.

FAQ sobre sistemas de archivos

- **¿Puedo ejecutar uno o más nodos del clúster como servidores NFS de alta disponibilidad con otros nodos de clúster como clientes?**

  No, no haga un montaje de bucle cerrado.

- **¿Puedo usar un sistema de archivos del clúster para aplicaciones que no estén bajo el control del Gestor de grupos de recursos?**

  Sí. Sin embargo, sin el control de RGM es necesario reiniciar las aplicaciones manualmente después que falle el nodo en el que se están ejecutando.

- **¿Deben tener todos los sistemas de archivo del clúster un punto de montaje debajo del directorio /global?**

  No. Sin embargo, si se sitúan los sistemas de archivos clúster bajo el mismo punto de montaje, como `/global`, se consigue una mejor organización y gestión de estos sistemas de archivos.

- **¿Cuáles son las diferencias entre usar el sistema de archivos del clúster y exportar sistemas de archivos NFS?**

  Existen varias diferencias:

  1. El sistema de archivos del clúster admite dispositivos globales. NFS no admite acceso remoto a los dispositivos.
  2. El sistema de archivos del clúster dispone de un espacio de nombres global. Sólo es necesario un comando de montaje. Con NFS, debe montarse el sistema de archivos en cada uno de los nodos.
  3. El sistema de archivos del clúster almacena en antememoria los archivos en más casos que NFS. Por ejemplo, cuando se accede a un archivo desde varios nodos, para lectura, escritura, bloqueo de archivo, E/S asíncrona.
  4. El sistema de archivos del clúster admite la recuperación de fallos si uno de los servidores falla. NFS admite varios servidores, pero la recuperación de fallos sólo es posible para sistemas de archivos de sólo lectura.

5. El sistema de archivos del clúster está creado para aprovechar futuras interconexiones rápidas de clúster que ofrezcan DMA remotas y funciones de copia cero.
6. Si se cambian los atributos de un archivo (mediante `chmod(1M)`, por ejemplo) en un sistema de archivos del clúster, el cambio se muestra de forma inmediata en todos los nodos. Con un sistema de archivos NFS exportado, esto puede llevar mucho más tiempo.

- **El sistema de archivos /global/.devices/node@<IDnodo> aparece en los clústers de mi nodo. ¿Puedo usar este sistema de archivos para almacenar datos que deseo que estén altamente disponibles y sean globales?**

  Estos sistemas de archivos almacenan los espacios de nombre de dispositivo global. No están pensados para un uso general. Aunque sean globales, nunca se accede a ellos de forma global (cada nodo sólo accede a su propio espacio de nombres de dispositivo global). Si un nodo está caído, el resto no puede acceder al espacio de nombres del nodo en cuestión. Estos sistemas de archivos no son de alta disponibilidad. No deberían usarse para almacenar datos que hayan de estar globalmente accesibles o altamente disponibles.

FAQ de gestión de volúmenes

- **¿Necesito duplicar todos los dispositivos de disco?**

  Para que un dispositivo de disco se considere de alta disponibilidad, debe estar duplicado o usar hardware RAID-5. Todos los servicios de datos deben usar dispositivos de disco de alta disponibilidad o sistemas de archivos del clúster montados en dispositivos de disco de alta disponibilidad. Estas configuraciones pueden tolerar fallos de disco puntuales.

- **¿Puedo usar un gestor de volúmenes para los discos locales (disco de arranque) y otro distinto para los discos multisistema?**

  Esta configuración se admite con el software Gestor de volúmenes de Solaris cuando gestiona los discos locales y VERITAS Volume Manager con los discos multisistema. No se admite ninguna otra combinación.

FAQ de servicios de datos

- **¿Qué servicios de datos SunPlex están disponibles?**

  La lista de servicios de datos admitidos se incluye en el manual *Sun Cluster 3.1: Notas sobre la versión*.

- **¿Qué versiones de aplicaciones admiten los servicios de datos de SunPlex?**

  La lista de las versiones de aplicaciones admitidas se incluye en el *Sun Cluster 3.1: Notas sobre la versión*.

- **¿Puedo escribir mi propio servicio de datos?**

  Sí. Para obtener más información, consulte la documentación de *Sun Cluster 3.1: Guía del desarrollador de los servicios de datos* y de Data Service Enabling Technologies que se incluía con la API de desarrollo de las bibliotecas del servicio de datos.

- **Al crear recursos de red, ¿debo especificar direcciones IP numéricas o nombres de sistema?**

  El método preferido para especificar recursos de red es usar el nombre de sistema de UNIX en lugar de la dirección IP numérica.

- **Al crear recursos de red, ¿cuál es la diferencia entre usar un nombre de sistema lógico (un recurso LogicalHostname) o una dirección compartida (un recurso SharedAddress)?**

  Excepto en el caso de Sun Cluster HA para NFS, siempre que en la documentación se habla del uso de un recurso `LogicalHostname` de un grupo de recursos de modo `Failover`, en su lugar se puede usar un recurso `SharedAddress` o `LogicalHostname`. El uso de un recurso `SharedAddress` produce una sobrecarga adicional porque el software de red del clúster está configurado para `SharedAddress` pero no para `LogicalHostname`.

  Existe una ventaja respecto a usar `SharedAddress` cuando que se está configurando servicios de datos escalables y a prueba de fallos y se desea que los clientes puedan acceder a ambos servicios usando el mismo nombre de sistema. En este caso, los recursos `SharedAddress` junto con el recurso de la aplicación de recuperación de fallos están contenidos en un grupo de recursos, mientras que el recurso del servicio escalable está contenido en un grupo de recursos separado y configurado para usar `SharedAddress`. Tanto los servicios escalables como los a prueba de fallos pueden usar después el mismo conjunto de nombres de sistema/direcciones que estén configurados en el recurso `SharedAddress`.

FAQ de redes públicas

- **¿Qué adaptadores de red pública admite el sistema SunPlex?**

  Actualmente el sistema SunPlex admite adaptadores de red pública Ethernet (10/100BASE-T y 1000BASE-SX Gb). Debido a que en el futuro podrían admitirse interfaces nuevas, compruebe con el representante de ventas de Sun cuál es la información más actual.

- **¿Cuál es el papel de la dirección MAC en la recuperación de fallos?**

Cuando se produce una recuperación de fallos, se generan nuevos paquetes de protocolo de resolución de direcciones (ARP) y se envían a todo el mundo. Estos paquetes ARP contienen la nueva dirección MAC (del nuevo adaptador físico al que se ha transferido el nodo) y la dirección IP antigua. Cuando otra máquina de la red recibe uno de estos paquetes, borra la correlación MAC-IP antigua de la memoria intermedia ARP y usa la nueva.

- **¿El sistema SunPlex admite el valor local-mac-address?=true de la PROM OpenBoot™ (OBP) para un adaptador de sistema?**

  Sí. Además, la Ruta múltiple de red IP requiere que `local-mac-address?` esté configurada como `true`.

- **¿Cuánto retraso puede esperarse cuando Ruta múltiple de red IP lleva a cabo un cambio entre adaptadores?**

  El retraso podría ser de varios minutos. Esto es debido a que cuando se lleva a cabo el cambio de Ruta múltiple de red IP, implica enviar una ARP innecesaria. Sin embargo, no hay garantías de que el encaminador entre el cliente y el clúster utilizará la ARP innecesaria. Así que, hasta que no se supere el tiempo de espera excedido de la entrada de memoria intermedia ARP para esta dirección IP en el encaminador, es posible que utilice la dirección MAC falsa.

- **¿Con qué velocidad se detecta el fallo de un adaptador de red?**

  El tiempo de detección de fallos predeterminado es de 10 segundos. Este algoritmo intenta cumplir el tiempo de detección de fallos predeterminado, pero el real depende de la carga de la red.

FAQ sobre pertenencia al clúster

- **¿Tienen que tener la misma contraseña de superusuario todos los miembros del clúster?**

  No es necesario que el superusuario comparta la misma contraseña en todos los miembros del clúster. Sin embargo, esta práctica simplifica mucho la administración del clúster.

- **¿Es importante el orden en el que arrancan los nodos?**

  En la mayoría de casos, no. Sin embargo, el orden de arranque es importante para evitar la amnesia (consulte «Quórum y dispositivos del quórum» en la página 52 para obtener detalles sobre la amnesia). Por ejemplo, si el nodo dos era el propietario del dispositivo del quórum y el nodo uno estaba caído, y después se desactiva el nodo dos, hay que volver a activar éste antes que aquél. Esto evita que accidentalmente se active un nodo con información de configuración del clúster anticuada.

- **¿Es necesario realizar una duplicación de los discos locales en un nodo del clúster?**

Sí. Aunque esta duplicación no es un requisito, con ella se asegura que un fallo de disco sin duplicación haga inoperativo el nodo. El inconveniente de realizar una duplicación de los discos locales de un nodo del clúster es que complica la administración del sistema.

- **¿Cuáles son los inconvenientes de realizar copias de respaldo de los miembros del clúster?**

  En un clúster se pueden usar varios métodos de copia de respaldo. Un método es tener un nodo como respaldo con una unidad de cinta/biblioteca conectada. A continuación se debe usar el sistema de archivos del clúster para realizar la copia de seguridad de los datos. No conecte este nodo a los discos compartidos.

  Consulte el manual *Sun Cluster 3.1: Guía de administración del sistema* para obtener información adicional sobre los procedimientos de copia de seguridad y restauración.

- **¿Cuándo está un nodo en el suficiente buen estado como para usarlo como secundario?**

  Después de un arranque, un nodo está en el suficiente buen estado como para ser secundario cuando éste muestra el indicador de inicio de sesión.

FAQ de almacenamiento del clúster

- **¿Qué hace que el almacenamiento multisistema sea de alta disponibilidad?**

  El almacenamiento multisistema es de alta disponibilidad porque puede sobrevivir a la pérdida de un disco, gracias a la duplicación (o debido a controladores RAID-5 basados en el hardware). Como los dispositivos de almacenamiento multisistema tienen más de una conexión de sistema, también pueden resistir la pérdida de un nodo al cual estén conectados. Además, las rutas redundantes desde cada nodo al almacenamiento conectado ofrecen tolerancia para el fallo de un adaptador del bus del sistema, de un cable o del controlador de disco.

FAQ de interconexión del clúster

- **¿Qué interconexiones del clúster admite el sistema SunPlex?**

  Actualmente, el sistema SunPlex admite interconexiones del clúster de interfaz de red Ethernet (100BASE-T Fast Ethernet y 1000BASE-SX Gb) y SCI.

- **¿Cuál es la diferencia entre un “cable” y una “ruta” de transporte?**

Los cables de transporte del clúster se configuran mediante los adaptadores de transporte y los conmutadores. Los cables unen adaptadores y conmutadores según cada componente. El gestor de la topología del clúster usa los cables disponibles para crear las rutas de transporte entre los extremos de los nodos. Un cable no se correlaciona directamente con una ruta de transporte.

Los cables son "habilitados" e "inhabilitados" por el administrador estáticamente. Los cables tienen un "estado," (habilitado o inhabilitado) pero no un "estatus." Si se inhabilita un cable, es como si estuviera sin configurar. Los cables que están inhabilitados no pueden usarse como rutas de transporte. No han sido sondeados y por tanto no es posible conocer su estatus. El estado de un cable puede verse con `scconf - p`.

El gestor de topología del clúster establece dinámicamente las rutas de transporte. El "estatus" de una ruta de transporte viene determinado por el gestor de topologías. Una ruta puede tener el estatus de "en línea" o "fuera de línea." El estatus de una ruta de transporte puede visualizarse con `scstat(1M)`.

Considere el ejemplo siguiente de clúster de dos nodos con cuatro cables.

```
node1:adapter0     to switch1, port0
node1:adapter1     to switch2, port0
node2:adapter0     to switch1, port1
node2:adapter1     to switch2, port1
```

A partir de estos cuatro cables se pueden formar dos posibles rutas de transporte.

```
node1:adapter0    to node2:adapter0
node2:adapter1    to node2:adapter1
```

FAQ sobre sistemas cliente

- **¿Es necesario considerar alguna necesidad o restricción de cliente para usarla con un clúster?**

  Los sistemas cliente se conectan al clúster como lo harían con cualquier otro servidor. En algunos casos, dependiendo de la aplicación del servicio de datos, es posible que necesite instalar software de lado del cliente o llevar a cabo otros cambios de configuración para que el cliente pueda conectarse a la aplicación del servicio de datos. Consulte los capítulos de *Sun Cluster 3.1 Data Services Installation and Configuration Guide* relacionados con este tema para obtener más información sobre los requisitos de la configuración del lado del cliente.

FAQ de consola de administración

- **¿Requiere el sistema SunPlex una consola de administración?**

  Sí.

- **¿Tiene que tratarse de una consola de administración exclusiva para el clúster? ¿O puede usarse para otras tareas?**

  El sistema SunPlex no requiere el uso de una consola de administración exclusiva, pero si se utiliza una de este tipo se obtienen estas ventajas:

  - Permite la gestión centralizada del clúster ya que agrupa herramientas de consola y de gestión en la misma máquina
  - Ofrece al proveedor de servicio de hardware una resolución de problemas potencialmente más rápida

- **¿Es necesario que la consola de administración se encuentre "cerca" del propio clúster, por ejemplo en la misma habitación?**

  Compruébelo con su proveedor de servicio de hardware. Es posible que el proveedor le requiera que la consola se sitúe muy cerca del clúster en sí mismo. No existe ninguna razón técnica por la que la consola deba estar situada en la misma habitación.

- **¿Puede servirse a más de un clúster desde la misma consola de administración, siempre que se cumplan los requisitos de distancia?**

  Sí. Se pueden controlar varios clústers desde una misma consola de administración. También puede compartirse un concentrador de terminal simple entre varios clústers.

FAQ sobre concentrador de terminal y procesador de servicio del sistema

- **¿Requiere el sistema SunPlex un concentrador de terminal?**

  Todas las versiones del software empezando por Sun Cluster 3.0 no requieren la ejecución de ningún concentrador de terminal. Al contrario de lo que ocurre con el producto Sun Cluster 2.2, que requería un concentrador de terminal para el aislamiento de fallos, los productos posteriores no dependen el concentrador de terminal.

- **Veo que la mayoría de servidores de SunPlex usan concentradores de terminal, pero el servidor Sun Enterprise E10000 no. ¿Por qué?**

El concentrador de terminal es en la práctica un conversor serie para Ethernet para la mayoría de los servidores. Su puerto de consola es un puerto serie. El servidor Sun Enterprise E10000 no dispone de consola serie. El procesador de servicio del sistema (SSP) es la consola, a través de Ethernet o del puerto `jtag`. Para el servidor Sun Enterprise E10000, siempre se usa SSP para consolas.

- **¿Cuáles son las ventajas de usar un concentrador de terminal?**

  Usar un concentrador de terminal ofrece acceso a nivel de consola a todos los nodos desde una estación de trabajo remota de la red, incluso cuando el nodo está en la PROM OpenBoot (OBP).

- **Si utilizo un concentrador de terminal no admitido por Sun, ¿qué necesito saber para certificar el que deseo utilizar?**

  La diferencia principal entre el concentrador de terminal que admite Sun y los otros dispositivos de consola es que el de Sun dispone de firmware especial que evita que éste envíe un carácter de interrupción a la consola cuando arranca. Tenga en cuenta que si dispone de un dispositivo de consola que envía a la consola caracteres de interrupción o una señal que pueda interpretarse como tales, el nodo se apaga.

- **¿Se puede liberar un puerto bloqueado en el concentrador de terminal que admite Sun sin volverlo a arrancar?**

  Sí. Anote el número de puerto que necesita restaurarse y haga lo siguiente:

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

  Consulte el manual *Sun Cluster 3.1: Guía de administración del sistema* para obtener más información sobre la configuración y administración del concentrador de terminal que admite Sun.

- **¿Qué ocurre si falla el propio concentrador de terminal?¿Es necesario tener otro de recambio?**

  No. No se pierde ninguna disponibilidad del clúster si el concentrador de terminal falla. Se pierde la posibilidad de conectarse a las consolas del nodo hasta que el concentrador vuelva a estar operativo.

- **Si se usa un concentrador de terminal, ¿qué ocurre con la seguridad?**

  En general, el concentrador de terminal está conectado a una pequeña red que usan los administradores de sistema, no una red que se use para otros accesos de clientes. La seguridad se puede controlar limitando el acceso a esa red en particular.

- **¿Cómo se utiliza la reconfiguración dinámica con una cinta o unidad de disco?**

- Determine si el disco o la unidad de cinta forma parte de un grupo de dispositivos activo. Si la unidad no forma parte de un grupo de dispositivos activos, puede llevar a cabo la operación de extracción DR en él.
- Si la operación de extracción de placa DR pudiera afectar a un disco o unidad de cinta activos, el sistema rechazará la operación e identificará las unidades que se verían afectadas por la operación. Si la unidad forma parte de un grupo de dispositivos activos, vaya a «Consideraciones de la agrupación DR para unidades de disco y cinta» en la página 84.
- Determine si la unidad es un componente del nodo primario o del secundario. Si la unidad es un componente del nodo secundario, puede llevar a cabo la operación de extracción DR en el mismo.
- Si la unidad es un componente del nodo primario, es necesario intercambiar los nodos primario y secundario antes de realizar la operación de extracción DR en el dispositivo.

Precaución – Si el nodo principal falla durante una operación de DR en un nodo secundario, la disponibilidad del clúster queda afectada. El nodo primario no tiene dónde transferirse hasta que se proporcione un nodo secundario nuevo.

Glosario

Este glosario de términos se utiliza en la documentación de SunPlex 3.1.

A

adaptador de transporte del clúster El adaptador de red que reside en un nodo y lo conecta a la interconexión del clúster. Consulte también “interconexión del clúster.”

amnesia Una condición en la que un clúster se reinicia después de un apagado con datos de configuración de clúster obsoletos (CCR). Por ejemplo, en un clúster de dos nodos con solo un nodo operativo, si se producen cambios de configuración del clúster en el nodo 1, los CCR del nodo 2 se convierten en obsoletos. Si el clúster se apaga y después se reinicia en el nodo 2, se produce una condición de amnesia debido a la obsolescencia de los CCR del nodo 2.

API de gestión de recursos (RMAPI) Interfaz de programación de aplicaciones dentro de un sistema SunPlex que hace a la aplicación altamente disponible en un entorno del clúster.

C

cables de transporte del clúster La conexión de red que conecta los distintos extremos. Una conexión entre adaptadores y tomas de transporte del clúster o entre dos adaptadores de transporte del clúster. Consulte también “interconexión del clúster.”

clúster Dos o más nodos o dominios interconectados que comparten un sistema de archivos del clúster y están configurados juntos para ejecutar recursos a prueba de fallos, paralelos o escalables.

concentrador de terminales En configuraciones que no son Enterprise 10000, dispositivo que es externo al clúster y se utiliza específicamente para comunicarse con miembros del clúster.

conjunto de discos Consulte “grupo de dispositivos”.

conmutación La transferencia ordenada de un grupo de recursos o dispositivos de un maestro (nodo) de un clúster a otro (o a varios maestros, si los grupos de recursos están configurados para varios primarios). Los cambios los inician los administradores mediante el comando `scswitch(1M)`.

consola de administración Una estación de trabajo que se utiliza para ejecutar software de administración del clúster.

control Consulte “recuperación de fallos.”

controlador DID Un controlador implementado por el software de Sun Cluster que se usa para ofrecer un espacio de nombres uniforme en todo el clúster. Consulte también “nombre DID”.

coubicación La propiedad de estar en el mismo nodo. Este concepto se usa durante la configuración del clúster para mejorar el rendimiento.

D

depósito de configuración del clúster (CCR) Un almacén de datos replicados de alta disponibilidad que utiliza el software Sun Cluster para almacenar de forma persistente información de configuración del clúster.

disco local Disco que es físicamente privado para un nodo del clúster dado.

disco multisistema Disco que está conectado físicamente a varios nodos.

dispositivo del quórum Disco compartido por dos o más nodos que contribuye con votos usados para establecer el quórum para que se ejecute el clúster. El clúster 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.

dispositivo global Dispositivo que está accesible desde todos los miembros del clúster, como discos, CD-ROM y cintas.

E

equilibrio de cargas Sólo se aplica a los servicios escalables. Proceso de distribuir la carga de la aplicación entre los nodos del clúster para que se sirva a las peticiones de los clientes de forma puntual. Consulte «Servicios de datos escalables» en la página 61 para obtener más detalles.

espacio de nombres de dispositivos globales Un espacio de nombres que contiene los nombres lógicos del clúster para dispositivos globales. Los dispositivos locales del entorno Solaris se definen en los directorios `/dev/dsk`, `/dev/rdsk` y `/dev/rmt`. El espacio de nombres de dispositivos globales define éstos en los directorios `/dev/global/dsk`, `/dev/global/rdsk` y `/dev/global/rmt`.

esquizofrenia Condición por la que un clúster se divide en varias particiones, sin que ninguna de ellas tenga conocimiento de las otras.

estado de grupo de recursos Estado del grupo de recursos en cualquier nodo dado.

estado de recurso Condición del recurso de acuerdo con la información del supervisor de fallos.

estado de recursos Estado de un recurso del Gestor de grupos de recursos en un nodo dado.

evento Un cambio en el estado, papel, gravedad o descripción de un objeto gestionado.

extremo Un puerto físico de un adaptador o toma de transporte del clúster.

G

Gestor de bloqueo distribuido (DLM) El software de bloqueo que se usa en un entorno de disco compartido de Oracle Parallel Server (OPS). DLM permite que los procesos Oracle que se ejecutan en nodos distintos sincronicen el acceso a las bases de datos. DLM está diseñado para alta disponibilidad. Si un proceso o nodo cae, el resto de nodos no tienen que apagarse y reiniciarse. Para recuperarse de este fallo, se realiza una reconfiguración rápida del DLM.

Gestor de grupos de recursos (RGM) Característica de software que se usa para hacer los recursos del clúster altamente disponibles y escalables iniciándolos y parándolos

	automáticamente en nodos de clúster seleccionados. RGM actúa según políticas preconfiguradas, en caso de fallos o reinicio de hardware o software.
gestor de volúmenes	Producto de software que ofrece fiabilidad de datos mediante división, concatenación, duplicación y crecimiento dinámico de metadispositivos o volúmenes.
Gestor de volúmenes de Solaris	Gestor de volúmenes usado por el sistema SunPlex. Consulte también "gestor de volúmenes".
grupo de discos	Consulte "grupo de dispositivos".
grupo de dispositivos	Un grupo definido por el usuario de recursos de dispositivos, como discos, que pueden tomarse de distintos nodos en una configuración del clúster AD. Este grupo puede incluir recursos de dispositivos de discos, conjuntos de discos del Gestor de volúmenes de Solaris y grupos de discos VERITAS Volume Manager.
grupo de dispositivos de disco	Consulte "grupo de dispositivos".
grupo de recursos	Colección de recursos gestionados por RGM como unidad. Cada recurso que va a gestionar RGM debe configurarse en un grupo de recursos. Normalmente los recursos relacionados e interdependientes están agrupados.
grupo de ruta múltiple de red IP	Conjunto de uno o más adaptadores de red del mismo nodo y subred configurados para respaldarse entre sí en caso de fallo del adaptador.
grupo de seguridad	Consulte "grupo de ruta múltiple de red IP".

I

id de dispositivo	Un mecanismo de identificación de los dispositivos que están disponibles a través de Solaris. Los id de dispositivo se describen en la página de comando man `devid_get(3DEVID)`. El controlador DID de Sun Cluster usa los id de dispositivo para determinar la correlación entre los nombres lógicos de Solaris de los distintos nodos del clúster. El controlador DID sondea todos los dispositivos para averiguar su id de dispositivo. Si éste coincide con el de otro que haya en el clúster, ambos reciben el mismo nombre DID. Si el id de dispositivo no se ha visto en el clúster anteriormente, se le asigna un nuevo nombre DID. Consulte también "nombre lógico de Solaris" y "controlador DID".

instancia Consulte “invocación de recursos”.

instancia paralela de servicio Una instancia de un tipo de recurso paralelo que se ejecuta en un nodo individual.

interconexión del clúster La infraestructura de hardware de la red que incluye cables, tomas y adaptadores de transporte del clúster. Sun Cluster y el software del servicio de datos usan esta infraestructura para la comunicación interna del clúster.

Interfaz coherente escalable (SCI) Hardware de interconexión de alta velocidad que se usa como interconexión del clúster.

interfaz global Una interfaz de red global que aloja físicamente direcciones compartidas. Consulte también “dirección compartida”.

interfaz lógica de red En la arquitectura de Internet, un sistema puede tener una o mas direcciones IP. El software de Sun Cluster configura interfaces de red lógicas adicionales para establecer una correlación entre varias interfaces de red lógicas y una única interfaz de red física. Cada interfaz lógica de red dispone de una única dirección IP. Esta correlación permite a una misma interfaz de red física que responda a varias direcciones IP distintas, así como que la dirección IP se traslade de un miembro del clúster a otro en caso de toma de control o conmutación sin requerir interfaces de hardware adicionales.

invocación de recursos Instancia de un tipo de recurso que se ejecuta en un nodo. Concepto abstracto que representa un recurso que se ha iniciado en el nodo.

IPv4 Internet Protocol, versión 4. Algunas veces denominado IP. Esta versión admite un espacio de direcciones de 32 bits.

L

latido Mensaje periódico enviado a todas las rutas de transporte de interconexión del clúster disponibles. Si falta el latido después de un intervalo y número especificados de intentos podría activar una recuperación de fallos interna de la comunicación de transporte a otra ruta. El fallo de todas las rutas a un miembro del clúster produce que CMM reevalúe el quórum del clúster.

M

maestro Consulte “primario”.

maestro potencial Consulte “principal potencial”.

maestro predeterminado El miembro de clúster predeterminado en el que se pone en línea un tipo de recurso de recuperación de fallos.

miembro del clúster Un miembro activo de la versión actual del clúster que es capaz de compartir recursos con otros miembros del clúster y de ofrecer servicios a otros miembros del clúster y a clientes de éste. Consulte también “nodo del clúster.”

modernización progresiva En una configuración de Sun Cluster, modernización que se lleva a cabo secuencialmente en un nodo del clúster cada vez. Durante una modernización progresiva, el clúster permanece en producción y los servicios continúan ejecutándose en los demás nodos.

modo sin clúster Estado resultante conseguido al arrancar un miembro del clúster con la opción de arranque `-x`. En este estado el nodo ya no es miembro del clúster, pero sigue siendo nodo del clúster. Consulte también “miembro del clúster” y “nodo del clúster”.

N

nodo Máquina o dominio físicos (en el servidor Sun Enterprise E10000) que puede formar parte de un sistema SunPlex. También se denomina “sistema.”

nodo de interfaz global (GIF) Nodo que aloja una interfaz global.

nodo del clúster Un nodo que está configurado para ser un miembro del clúster. Un nodo del clúster podría ser o no miembro actual del clúster. Consulte también “miembro del clúster”.

nodo GIF Consulte ”nodo de interfaz global”.

nombre de sistema principal Nombre de un nodo de la red pública primaria. Siempre es el nombre del nodo especificado en `/etc/nodename`. Consulte también “nombre de sistema secundario”.

nombre de sistema privado Alias de nombre del sistema usado para comunicarse con un nodo a través de la interconexión del clúster.

nombre de sistema secundario Nombre usado para acceder a un nodo en una red secundaria pública. Consulte también “nombre de sistema principal”.

nombre DID Nombre que se usa para identificar dispositivos globales en un sistema SunPlex. Es un identificador de clúster con una relación de tipo uno a uno o uno a varios con nombres lógicos de Solaris. Toma la forma d*X*s*Y*, donde *X* es un entero e *Y* es el nombre de segmento. Consulte también “nombre lógico de Solaris ”.

nombre físico de Solaris Nombres que los controladores de dispositivos dan a los dispositivos en Solaris. Aparece en las máquinas Solaris como nombre de ruta en el árbol `/devices`. Por ejemplo, un disco SCSI típico tiene un nombre físico de Solaris parecido a:`/devices/sbus@1f,0/SUNW,fas@e,8800000/sd@6,0:c,raw`

Consulte también “nombre lógico de Solaris ”.

nombre lógico de Solaris Nombres usados normalmente para gestionar dispositivos de Solaris. En el caso de discos normalmente son parecidos a: `/dev/rdsk/c0t2d0s2`. Para cada uno de estos nombres lógicos de Solaris, existe un nombre de dispositivo físico de Solaris subyacente. Consulte también “nombre DID” y “nombre físico de Solaris”.

P

política de equilibrio de cargas Sólo se aplica a los servicios escalables. La forma preferida en la que las aplicaciones solicitan la carga es la distribución entre nodos. Consulte «Servicios de datos escalables» en la página 61 para obtener más detalles.

primario Nodo en el que que está en línea un grupo de recursos o dispositivos. Es decir, un primario es un nodo que está alojando o implementando en este momento el servicio asociado con el recurso. Consutle también “secundario.”

principal potencial Miembro del clúster que es capaz de gestionar un tipo de recurso de recuperación de fallos si el nodo primario falla. Consulte también “maestro predeterminado”.

Procesador de servicio del sistema (SSP) En las configuraciones Enterprise 10000, un dispositivo externo al clúster que se usa específicamente para comunicarse con miembros del clúster.

propiedad del tipo de recurso Un par de valores de clave que RGM almacena como parte del tipo de recurso y que se usa para describir y gestionar recursos de un tipo dado.

punto de comprobación La notificación que envía un nodo primario a uno secundario para mantener el estado del software sincronizado entre ambos. Consulte también “primario” y “secundario”.

R

recuperación de fallos La reubicación automática de un grupo de recursos o dispositivos desde un nodo primario actual a uno nuevo después de que se haya producido un fallo.

recuperación rápida El apagado y la retirada ordenados de un nodo fallido del clúster antes de que su funcionamiento potencialmente incorrecto se demuestre peligrosos.

recurso Instancia de un tipo de recurso. Pueden existir muchos recursos del mismo tipo, cada uno con su propio nombre y conjunto de valores de propiedad, por lo que varias instancias de la aplicación subyacente podrían ejecutarse en el clúster.

recurso a prueba de fallos Un recurso, cada uno de cuyos recursos sólo puede gestionarlos un solo nodo simultáneamente. Consulte también “recurso de instancia única” y “recurso escalable”.

recurso de dirección compartida Dirección de red que los servicios escalables que se ejecutan en los nodos del clúster pueden vincular para escalar éstos en los nodos. Un clúster puede tener varias direcciones compartidas y un servicio puede vincularse a varias direcciones compartidas.

recurso de dirección de red Consulte “recurso de red”.

recurso de nombre lógico de servidor Recurso que contiene una colección de nombres de sistema lógicos que representan direcciones de red. Estos recursos sólo puede gestionarlos un nodo cada vez. Consulte también “sistema lógico.”

recurso de red Recurso que contiene uno o más nombres de sistema lógicos o direcciones compartidas. Consulte también “recurso de nombre lógico de servidor” y “recurso de dirección compartida”.

recurso escalable Recurso que se ejecuta en varios nodos (una copia en cada uno) y usa la interconexión del clúster para dar la apariencia de un servicio único a los clientes remotos.

recurso genérico El daemon de aplicación y sus procesos subordinados puestos bajo el control del gestor de grupos de recursos como parte del tipo de recurso genérico.

recurso global Recurso de alta disponibilidad que se proporciona a nivel de núcleo del software de Sun Cluster. Los recursos globales pueden incluir discos (grupos de dispositivos AD), el sistema de archivos del clúster e interconexión de red global.

réplica de base de datos de metadispositivos (réplica) Base de datos almacenada en disco que registra la configuración y el estado de todos los metadispositivos y las condiciones de error. Esta información es importante para el correcto funcionamiento de los conjuntos de discos del Gestor de volúmenes de Solaris y se replica.

repuesto Nodo del clúster que está disponible para convertirse en secundario si se produce una recuperación de fallos. Consulte también “secundario”.

retroceso Consulte “retroceso automático”.

retroceso automático Proceso de volver un grupo de recursos o de dispositivos a su nodo primario después de que haya fallado y se haya reiniciado más tarde como miembro del clúster.

Ruta múltiple de red de protocolo de Internet (IP) Software que utiliza la supervisión y la recuperación de fallos para evitar la pérdida de disponibilidad del nodo debido a un único adaptador de red o fallo del cable. La recuperación de fallos por ruta múltiple de red IP usa conjuntos de adaptadores de red llamados grupos de ruta múltiple de red IP (grupos de ruta múltiple) para ofrecer conexiones redundantes entre un nodo del clúster y la red pública. Las posibilidades de supervisión y recuperación de fallos trabajan conjuntamente para asegurar la disponibilidad de los recursos. Consulte también “grupo de ruta múltiple de red IP”.

S

secundario Miembro del clúster que está disponible para grupos de dispositivos de disco maestros y grupos de recursos en caso de que el primario falle. Consulte también “primario”.

servicio de datos Una aplicación que se ha instrumentado para que se ejecute como recurso de alta disponibilidad bajo el control del gestor de grupo de recursos (RGM).

servicio de datos de alta disponibilidad (HA) Consulte “servicio de datos”.

servicio escalable Servicio de datos implementado que se ejecuta en varios nodos simultáneamente.

sistema de archivos del clúster Un servicio del clúster que ofrece acceso de alta disponibilidad a nivel de clúster a sistemas de archivo locales.

sistema de origen múltiple Sistema que está en más de una red pública.

sistema lógico Concepto de Sun Cluster 2.0 (mínimo) que incluye una aplicación, los conjuntos o grupos de discos en los que ésta reside y las direcciones de red usadas para acceder al clúster. Este concepto ya no existe en el sistema SunPlex. Consulte «Grupos de dispositivos de disco» en la página 40 y «Recursos: definición, grupos y tipos» en la página 69 para obtener una descripción de cómo se implementa este concepto ahora en el sistema SunPlex.

Sun Cluster (software) Parte del software del sistema SunPlex. (Consulte SunPlex).

SunPlex Sistema de hardware y software de Sun Cluster integrados que se usa para crear servicios de alta disponibilidad y escalabilidad.

supervisor de fallos El daemon de fallos y los programas que se usan para sondear las distintas partes de los servicios de datos y tomar las medidas necesarias. Consulte también “supervisor de recursos”.

supervisor de pertenencia al clúster (CMM) El software que mantiene un registro uniforme de los miembros del clúster. Esta información sobre pertenencia la usa el resto del software del clúster para decidir dónde situar los servicios de alta disponibilidad. CCM se asegura de que los miembros que no sean del clúster no puedan corromper datos y transmitir datos corruptos o incoherentes a los clientes.

supervisor de recursos Parte opcional de una implementación de un tipo de recurso que ejecuta periódicamente sondas de fallos en recursos para determinar si se están ejecutando correctamente y cuál es su rendimiento.

switchback Consulte “retroceso.”

T

tipo de recurso Nombre exclusivo dado a un servicio de datos, LogicalHostname u objeto de clúster SharedAddress. Los tipos de recurso del servicio de datos pueden ser a prueba de fallos o escalables. Consulte también “servicio de datos”, “recurso a prueba de fallos” y “recurso escalable”.

tipo de recurso de instancia única Recurso para el que como máximo puede haber un recurso activo en todo el clúster.

tipo de recurso genérico Plantilla para el servicio de datos. Tipo de recurso genérico que puede usarse para convertir una aplicación sencilla en un servicio de datos de recuperación de fallos (parar en un nodo, iniciar en otro). Este tipo no requiere programación de API SunPlex.

tipo de recurso paralelo Tipo de recurso, como una base de datos paralela, que se ha instrumentado para que se ejecute en un entorno de clúster para que pueda gestionarse simultáneamente desde varios (dos o más) nodos.

toma de transporte del clúster Un conmutador de hardware que se usa como parte de la interconexión del clúster. Consulte también “interconexión del clúster.”

V

VERITAS Volume Manager Gestor de volúmenes usado por el sistema SunPlex. Consulte también “gestor de volúmenes”.

Índice

D

E

F

G

H

T

U

V