MetroCluster

Enterprise applications

NetApp
February 11, 2026

Índice

MetroCluster

Recuperação de desastres com o MetroCluster

O MetroCluster é um recurso ONTAP que pode proteger seus bancos de dados Oracle com espelhamento síncrono de 0 RPO entre locais e escala para oferecer suporte a centenas de bancos de dados em um único sistema MetroCluster.

Também é simples de usar. O uso do MetroCluster não necessariamente adiciona ou altera quaisquer melhores pistas para operar aplicativos e bancos de dados empresariais.

As práticas recomendadas usuais ainda se aplicam. E, se suas necessidades exigirem apenas proteção de dados RPO de 0, essa necessidade será atendida com o MetroCluster. No entanto, a maioria dos clientes usa o MetroCluster não apenas para proteção de dados RPO igual a 0, mas também para aprimorar o rto durante cenários de desastre, bem como para fornecer failover transparente como parte das atividades de manutenção do local.

Arquitetura física

Entender como os bancos de dados Oracle operam em um ambiente MetroCluster requer alguma explicação do design físico de um sistema MetroCluster.

Esta documentação substitui o relatório técnico publicado anteriormente *TR-4592: Oracle no MetroCluster.*

O MetroCluster pode ser usado em 3 configurações diferentes

- Pares HA com conetividade IP
- Pares DE HA com conectividade FC
- Controladora única com conectividade FC

O termo 'conetividade' refere-se à conexão de cluster usada para replicação entre sites. Não se refere aos protocolos de host. Todos os protocolos do lado do host são suportados como de costume em uma configuração MetroCluster, independentemente do tipo de conexão usada para comunicação entre clusters.

IP MetroCluster

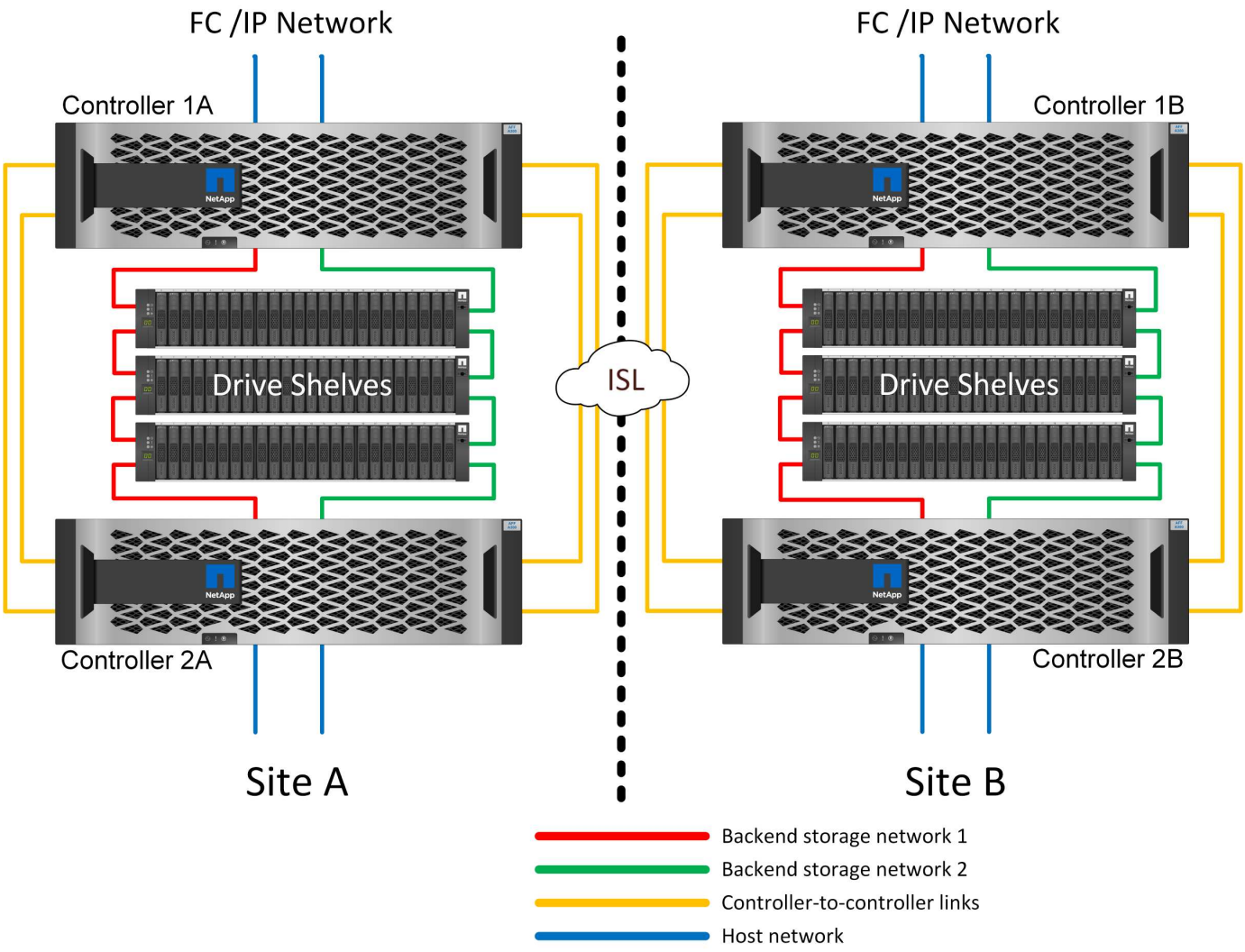
A configuração MetroCluster IP de par de HA usa dois ou quatro nós por local. Essa opção de configuração aumenta a complexidade e os custos em relação à opção de dois nós, mas oferece um benefício importante: A redundância intrasite. Uma simples falha do controlador não requer acesso aos dados na WAN. O acesso aos dados permanece local por meio do controlador local alternativo.

A maioria dos clientes está escolhendo a conetividade IP porque os requisitos de infraestrutura são mais simples. No passado, a conetividade entre locais de alta velocidade era geralmente mais fácil de provisionar usando switches FC e fibra escura, mas hoje em dia os circuitos IP de baixa latência e alta velocidade estão mais prontamente disponíveis.

A arquitetura também é mais simples porque as únicas conexões entre locais são para os controladores. No

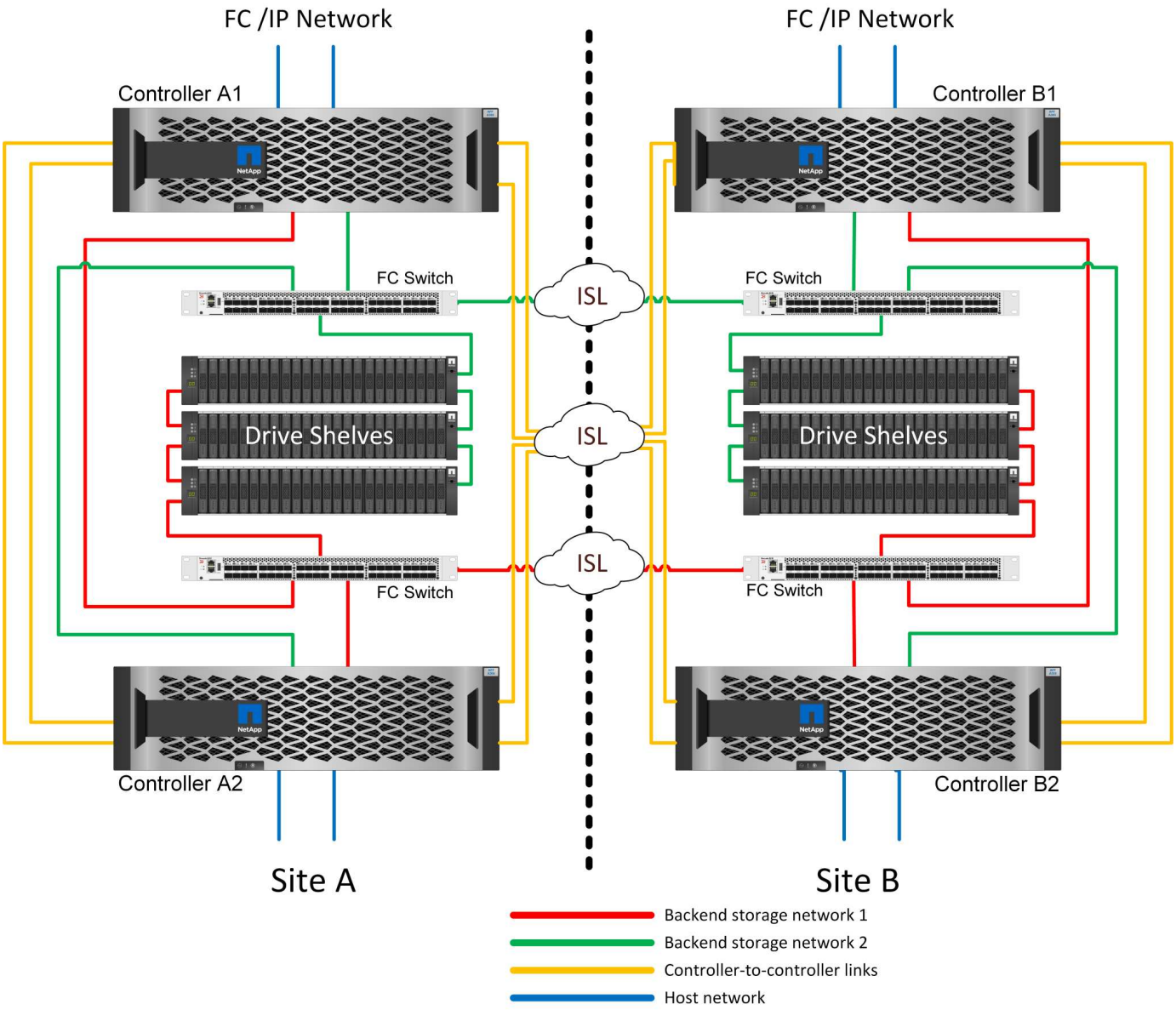
Metroclusters FC SAN conetados, um controlador grava diretamente nas unidades no local oposto e, portanto, requer conexões, switches e bridges SAN adicionais. Em contraste, um controlador em uma configuração IP grava nas unidades opostas através do controlador.

Para obter informações adicionais, consulte a documentação oficial do ONTAP e "Arquitetura e design da solução IP da MetroCluster".



MetroCluster conectados a FC de par HA SAN

A configuração de MetroCluster FC de par de HA usa dois ou quatro nós por local. Essa opção de configuração aumenta a complexidade e os custos em relação à opção de dois nós, mas oferece um benefício importante: A redundância intrasite. Uma simples falha do controlador não requer acesso aos dados na WAN. O acesso aos dados permanece local por meio do controlador local alternativo.

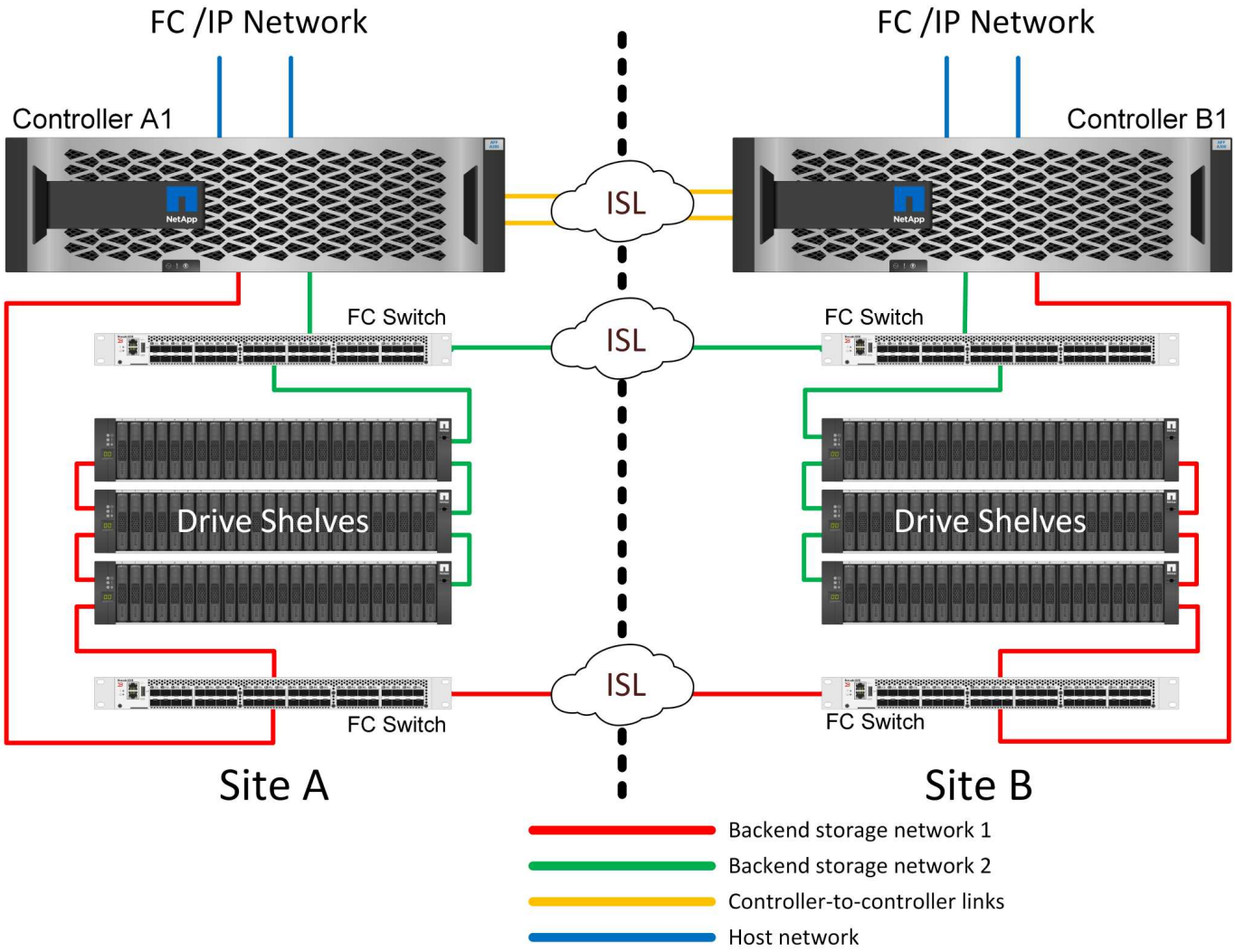


Algumas infraestruturas multisite não foram desenvolvidas para operações ativas-ativas, mas são usadas mais como local principal e local de recuperação de desastres. Nesta situação, uma opção MetroCluster de par de HA é geralmente preferível pelas seguintes razões:

- Embora um cluster de dois nós MetroCluster seja um sistema de HA, a falha inesperada de uma controladora ou a manutenção planejada exige que os serviços de dados fiquem online no local oposto. Se a conetividade de rede entre locais não puder suportar a largura de banda necessária, o desempenho é afetado. A única opção seria também falhar sobre os vários sistemas operacionais host e serviços associados ao site alternativo. O cluster de MetroCluster de par de HA elimina esse problema porque a perda de uma controladora resulta em failover simples no mesmo local.
- Algumas topologias de rede não são projetadas para acesso entre sites, mas usam sub-redes diferentes ou SANs FC isoladas. Nesses casos, o cluster MetroCluster de dois nós não funciona mais como um sistema de HA porque o controlador alternativo não pode fornecer dados aos servidores no local oposto. A opção MetroCluster de par de HA é necessária para fornecer redundância completa.
- Se uma infraestrutura de dois locais for vista como uma única infraestrutura altamente disponível, a configuração de dois nós do MetroCluster será adequada. No entanto, se o sistema precisar funcionar por um longo período de tempo após a falha do local, é preferível usar um par de HA porque ele continua fornecendo HA em um único local.

MetroCluster com conexão SAN FC de dois nós

A configuração do MetroCluster de dois nós usa apenas um nó por local. Esse design é mais simples do que a opção de par de HA, pois há menos componentes para configurar e manter. Ele também reduziu as demandas de infraestrutura em termos de cabeamento e switch FC. Finalmente, reduz custos.



O impactos óbvio desse projeto é que a falha do controlador em um único local significa que os dados estão disponíveis no local oposto. Esta restrição não é necessariamente um problema. Muitas empresas têm operações de data center multisite com redes estendidas, de alta velocidade e baixa latência, que funcionam essencialmente como uma única infraestrutura. Nesses casos, a versão de dois nós do MetroCluster é a configuração preferida. Sistemas de dois nós são usados atualmente em escala de petabyte por vários provedores de serviços.

Recursos de resiliência do MetroCluster

Não há pontos únicos de falha em uma solução MetroCluster:

- Cada controladora tem dois caminhos independentes para os compartimentos de unidades no local.
- Cada controladora tem dois caminhos independentes para o shelves de unidades no local remoto.
- Cada controlador tem dois caminhos independentes para os controladores no local oposto.

- Na configuração de par de HA, cada controladora tem dois caminhos para seu parceiro local.

Em resumo, qualquer componente na configuração pode ser removido sem comprometer a capacidade do MetroCluster de fornecer dados. A única diferença em termos de resiliência entre as duas opções é que a versão do par de HA ainda é um sistema de storage de HA geral após uma falha do local.

Arquitetura lógica

Entender como os bancos de dados Oracle operam em um ambiente MetroCluster o alsop requer alguma explicação da funcionalidade lógica de um sistema MetroCluster.

Proteção contra falha do local: NVRAM e MetroCluster

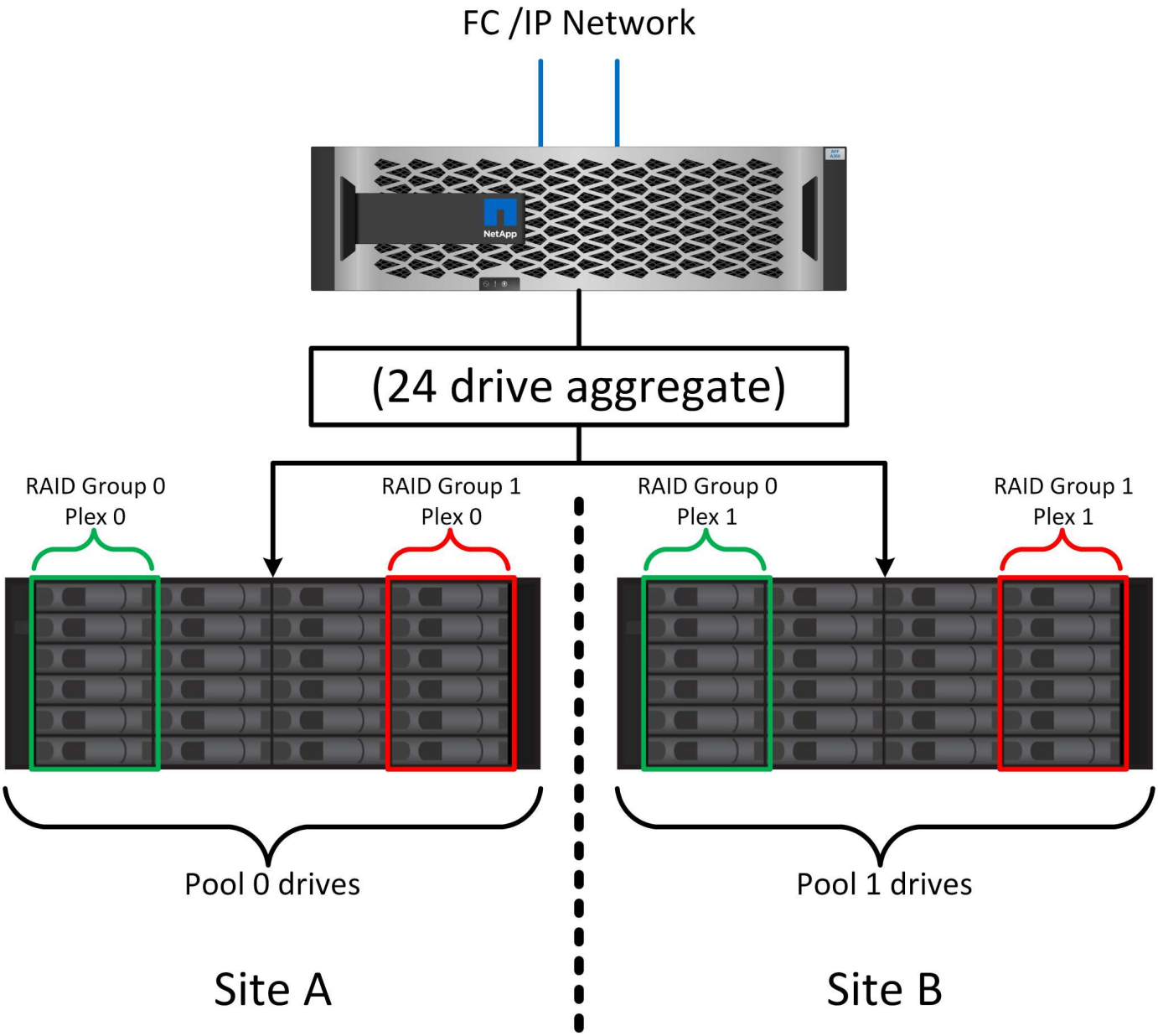
A MetroCluster estende a proteção de dados da NVRAM das seguintes maneiras:

- Em uma configuração de dois nós, os dados do NVRAM são replicados usando os ISLs (Inter-Switch Links) para o parceiro remoto.
- Em uma configuração de par de HA, os dados do NVRAM são replicados para o parceiro local e para um parceiro remoto.
- Uma gravação não é reconhecida até que seja replicada para todos os parceiros. Essa arquitetura protege a e/S em trânsito contra falhas do local replicando dados do NVRAM para um parceiro remoto. Este processo não está envolvido com a replicação de dados no nível da unidade. A controladora que detém os agregados é responsável pela replicação de dados por gravação em ambos os plexos no agregado, mas ainda deve haver proteção contra perda de e/S em trânsito em caso de perda do local. Os dados replicados do NVRAM só serão usados se um controlador do parceiro precisar assumir o controle de uma controladora com falha.

Proteção contra falhas no local e no compartimento: SyncMirror e plexos

O SyncMirror é uma tecnologia de espelhamento que aprimora, mas não substitui, o RAID DP ou o RAID-teC. Ele espelha o conteúdo de dois grupos RAID independentes. A configuração lógica é a seguinte:

1. As unidades são configuradas em dois pools com base no local. Um pool é composto por todas as unidades no local A, e o segundo pool é composto por todas as unidades no local B..
2. Um pool comum de armazenamento, conhecido como agregado, é criado com base em conjuntos espelhados de grupos RAID. Um número igual de unidades é extraído de cada local. Por exemplo, um agregado SyncMirror de 20 unidades seria composto por 10 unidades do local A e 10 unidades do local B..
3. Cada conjunto de unidades em um determinado local é configurado automaticamente como um ou mais grupos RAID DP ou RAID-teC totalmente redundantes, independentemente do uso do espelhamento. Esse uso de RAID por baixo do espelhamento fornece proteção de dados mesmo após a perda de um site.



A figura acima ilustra um exemplo de configuração do SyncMirror. Um agregado de 24 unidades foi criado na controladora com 12 unidades de um compartimento alocado no local A e 12 unidades de um compartimento alocado no local B. as unidades foram agrupadas em dois grupos RAID espelhados. RAID grupo 0 inclui um Plex de 6 unidades no local Um espelhado para um Plex de 6 unidades no local B. da mesma forma, RAID grupo 1 inclui um Plex de 6 unidades no local Um espelhado para um Plex de 6 unidades no local B.

O SyncMirror normalmente é usado para fornecer espelhamento remoto com sistemas MetroCluster, com uma cópia dos dados em cada local. Ocasionalmente, ele tem sido usado para fornecer um nível extra de redundância em um único sistema. Em particular, ele fornece redundância em nível de prateleira. Um compartimento de unidades já contém fontes de alimentação duplas e controladores e é, no geral, pouco mais do que chapas metálicas, mas em alguns casos, a proteção extra pode ser garantida. Por exemplo, um cliente da NetApp implantou o SyncMirror para uma plataforma móvel de análise em tempo real usada durante testes automotivos. O sistema foi separado em dois racks físicos fornecidos com alimentação de energia independente e sistemas UPS independentes.

Falha de redundância: NVFAIL

Como discutido anteriormente, uma gravação não é reconhecida até que ela tenha sido registrada no NVRAM local e no NVRAM em pelo menos um outro controlador. Essa abordagem garante que uma falha de hardware ou falha de energia não resulte na perda de e/S em trânsito Se o NVRAM local falhar ou a conectividade com outros nós falhar, os dados não serão mais espelhados.

Se o NVRAM local relatar um erro, o nó será encerrado. Esse desligamento resulta em failover para uma controladora de parceiro quando os pares de HA são usados. Com o MetroCluster, o comportamento depende da configuração geral escolhida, mas pode resultar em failover automático para a nota remota. Em qualquer caso, nenhum dado é perdido porque o controlador que está tendo a falha não reconheceu a operação de gravação.

Uma falha de conectividade local a local que bloqueia a replicação do NVRAM para nós remotos é uma situação mais complicada. As gravações não são mais replicadas nos nós remotos, criando uma possibilidade de perda de dados se ocorrer um erro catastrófico em um controlador. Mais importante ainda, tentar fazer failover para um nó diferente durante essas condições resulta em perda de dados.

O fator de controle é se o NVRAM está sincronizado. Se o NVRAM estiver sincronizado, o failover de nó para nó será seguro para prosseguir sem risco de perda de dados. Em uma configuração do MetroCluster, se o NVRAM e os plexos agregados subjacentes estiverem sincronizados, é seguro prosseguir com o switchover sem risco de perda de dados.

O ONTAP não permite um failover ou switchover quando os dados estão fora de sincronia, a menos que o failover ou switchover seja forçado. Forçar uma alteração de condições desta forma reconhece que os dados podem ser deixados para trás no controlador original e que a perda de dados é aceitável.

Bancos de dados e outros aplicativos ficam especialmente vulneráveis à corrupção se um failover ou switchover for forçado, porque eles mantêm caches internos maiores de dados no disco. Se ocorrer um failover forçado ou switchover, as alterações anteriormente confirmadas serão efetivamente descartadas. O conteúdo da matriz de armazenamento salta efetivamente para trás no tempo, e o estado do cache não reflete mais o estado dos dados no disco.

Para evitar essa situação, o ONTAP permite que os volumes sejam configurados para proteção especial contra falha do NVRAM. Quando acionado, esse mecanismo de proteção resulta em um volume entrando em um estado chamado NVFAIL. Esse estado resulta em erros de e/S que causam uma falha no aplicativo. Esta falha faz com que os aplicativos sejam desligados para que eles não usem dados obsoletos. Os dados não devem ser perdidos porque quaisquer dados de transação confirmados devem estar presentes nos logs. As próximas etapas usuais são para que um administrador desligue totalmente os hosts antes de colocar manualmente os LUNs e volumes novamente on-line. Embora essas etapas possam envolver algum trabalho, essa abordagem é a maneira mais segura de garantir a integridade dos dados. Nem todos os dados exigem essa proteção, e é por isso que o comportamento do NVFAIL pode ser configurado volume a volume.

Pares DE HA e MetroCluster

O MetroCluster está disponível em duas configurações: Dois nós e par de HA. A configuração de dois nós se comporta da mesma forma que um par de HA em relação ao NVRAM. Em caso de falha repentina, o nó do parceiro pode repetir dados do NVRAM para tornar as unidades consistentes e garantir que nenhuma gravação reconhecida tenha sido perdida.

A configuração de par de HA replica NVRAM também para o nó do parceiro local. Uma simples falha do controlador resulta em uma repetição do NVRAM no nó do parceiro, como é o caso de um par de HA autônomo sem MetroCluster. Em caso de perda súbita total do local, o local remoto também tem o NVRAM necessário para tornar as unidades consistentes e começar a fornecer dados.

Um aspeto importante do MetroCluster é que os nós remotos não têm acesso aos dados do parceiro em condições operacionais normais. Cada site funciona essencialmente como um sistema independente que pode assumir a personalidade do site oposto. Esse processo é conhecido como switchover e inclui um switchover planejado no qual as operações do local são migradas sem interrupções para o local oposto. Ele também inclui situações não planejadas em que um local é perdido e um switchover manual ou automático é necessário como parte da recuperação de desastres.

Comutação e comutação

Os termos switchover e switchback referem-se ao processo de transição de volumes entre controladores remotos em uma configuração do MetroCluster. Este processo aplica-se apenas aos nós remotos. Quando o MetroCluster é usado em uma configuração de quatro volumes, o failover de nó local é o mesmo processo de aquisição e giveback descrito anteriormente.

Comutação planejada e switchback

Um switchover planejado ou switchback é semelhante a um takeover ou giveback entre nós. O processo tem várias etapas e pode parecer exigir vários minutos, mas o que está realmente acontecendo é uma transição graciosa multifásica de recursos de armazenamento e rede. O momento em que as transferências de controle ocorrem muito mais rapidamente do que o tempo necessário para que o comando completo seja executado.

A principal diferença entre o takeover/giveback e o switchover/switchback está com o efeito na conectividade FC SAN. Com a takeover local/giveback, um host sofre a perda de todos os caminhos FC para o nó local e conta com o MPIO nativo para mudar para caminhos alternativos disponíveis. As portas não são realocadas. Com o switchover e o switchback, as portas de destino FC virtual nos controladores fazem a transição para o outro local. Eles efetivamente deixam de existir na SAN por um momento e, em seguida, reaparecem em um controlador alternativo.

Tempos limite do SyncMirror

O SyncMirror é uma tecnologia de espelhamento ONTAP que fornece proteção contra falhas nas shelves. Quando as gavetas são separadas à distância, o resultado é a proteção de dados remota.

O SyncMirror não fornece espelhamento síncrono universal. O resultado é uma melhor disponibilidade. Alguns sistemas de storage usam espelhamento constante de tudo ou nada, às vezes chamado de modo domino. Essa forma de espelhamento é limitada no aplicativo porque toda atividade de gravação deve cessar se a conexão com o local remoto for perdida. Caso contrário, uma escrita existiria em um site, mas não no outro. Normalmente, esses ambientes são configurados para colocar LUNs off-line se a conetividade site-a-site for perdida por mais de um curto período (como 30 segundos).

Este comportamento é desejável para um pequeno subconjunto de ambientes. No entanto, a maioria dos aplicativos exige uma solução que ofereça replicação síncrona garantida em condições operacionais normais, mas com a capacidade de suspender a replicação. Uma perda completa de conetividade local a local é frequentemente considerada uma situação de quase desastre. Normalmente, esses ambientes são mantidos on-line e fornecem dados até que a conectividade seja reparada ou uma decisão formal seja tomada para encerrar o ambiente para proteger os dados. Um requisito para o desligamento automático do aplicativo puramente por causa de falha de replicação remota é incomum.

O SyncMirror dá suporte aos requisitos de espelhamento síncrono com a flexibilidade de um tempo limite. Se a conetividade com o telecomando e/ou Plex for perdida, um temporizador de 30 segundos começa a contagem decrescente. Quando o contador atinge 0, o processamento de e/S de escrita é retomado utilizando os dados locais. A cópia remota dos dados é utilizável, mas fica congelada no tempo até que a conetividade seja restaurada. A ressincronização utiliza snapshots em nível agregado para retornar o sistema ao modo síncrono o mais rápido possível.

Notavelmente, em muitos casos, esse tipo de replicação universal do modo dominó tudo ou nada é melhor implementado na camada de aplicativo. Por exemplo, o Oracle DataGuard inclui o modo de proteção máximo, o que garante replicação de longa instância em todas as circunstâncias. Se o link de replicação falhar por um período que excede um tempo limite configurável, os bancos de dados serão desligados.

Switchover automático sem supervisão com MetroCluster conectado à malha

O switchover automático sem supervisão (AUSO) é um recurso de MetroCluster anexado a malha que fornece uma forma de HA entre os locais. Como discutido anteriormente, o MetroCluster está disponível em dois tipos: Um único controlador em cada local ou um par de HA em cada local. A principal vantagem da opção HA é que o desligamento planejado ou não planejado do controlador ainda permite que todas as I/o sejam locais. A vantagem da opção de nó único é reduzir os custos, a complexidade e a infraestrutura.

O principal valor do AUSO é melhorar os recursos de HA dos sistemas MetroCluster conectados a malha. Cada local monitora a integridade do local oposto e, se nenhum nó permanecer para fornecer dados, o AUSO resulta em switchover rápido. Essa abordagem é especialmente útil nas configurações do MetroCluster com apenas um nó único por local, pois aproxima a configuração de um par de HA em termos de disponibilidade.

A AUSO não pode oferecer monitoramento abrangente no nível de um par de HA. Um par de HA pode fornecer disponibilidade extremamente alta porque inclui dois cabos físicos redundantes para comunicação direta de nó a nó. Além disso, ambos os nós de um par de HA têm acesso ao mesmo conjunto de discos em loops redundantes, entregando outra rota para um nó monitorar a integridade de outro.

Os clusters do MetroCluster existem em locais para os quais a comunicação nó a nó e o acesso ao disco dependem da conectividade de rede local a local. A capacidade de monitorar o batimento cardíaco do restante do cluster é limitada. AUSO tem que discriminar entre uma situação em que o outro site está realmente inativo, em vez de indisponível devido a um problema de rede.

Como resultado, uma controladora em um par de HA pode solicitar um takeover se detetar uma falha na controladora que ocorreu por um motivo específico, como pânico do sistema. Ele também pode solicitar uma aquisição se houver uma perda completa de conetividade, às vezes conhecida como batimento cardíaco perdido.

Um sistema MetroCluster só pode efetuar uma mudança automática em segurança quando é detetada uma avaria específica no local original. Além disso, a controladora que assume a propriedade do sistema de storage deve ser capaz de garantir que os dados do disco e do NVRAM estejam sincronizados. O controlador não pode garantir a segurança de uma mudança apenas porque perdeu o Contato com o local de origem, que ainda poderia estar operacional. Para obter opções adicionais para automatizar um switchover, consulte as informações sobre a solução MetroCluster tiebreaker (MCTB) na próxima seção.

Desempate MetroCluster com MetroCluster conectado à malha

"Desempate de NetApp MetroCluster"O software pode ser executado em um terceiro local para monitorar a integridade do ambiente MetroCluster, enviar notificações e, opcionalmente, forçar um switchover em uma situação de desastre. Uma descrição completa do desempate pode ser encontrada no "Site de suporte da NetApp", mas o principal objetivo do desempate do MetroCluster é detetar a perda do local. Ele também deve discriminar entre a perda do local e a perda de conetividade. Por exemplo, o switchover não deve ocorrer porque o tiebreaker não conseguiu chegar ao local principal, e é por isso que o tiebreaker também monitora a capacidade do local remoto de entrar em Contato com o local principal.

O switchover automático com AUSO também é compatível com o MCTB. O AUSO reage muito rapidamente porque foi concebido para detetar eventos de falha específicos e, em seguida, invocar o switchover apenas quando os plexos NVRAM e SyncMirror estão em sincronia.

Em contraste, o desempate está localizado remotamente e, portanto, deve esperar que um temporizador

decorra antes de declarar um local morto. O tiebreaker eventualmente deteta o tipo de falha de controladora coberta pelo AUSO, mas, em geral, a AUSO já iniciou o switchover e possivelmente concluiu o switchover antes que o tiebreaker atue. O segundo comando de comutação resultante vindo do tiebreaker seria rejeitado.

> O software MCTB não verifica se o NVRAM estava e/ou os plexos estão em sincronia ao forçar um switchover. O switchover automático, se configurado, deve ser desativado durante atividades de manutenção que resultem na perda de sincronização para NVRAM ou SyncMirror plexes.

Além disso, o MCTB pode não resolver um desastre contínuo que leva à seguinte sequência de eventos:

1. A conetividade entre locais é interrompida durante mais de 30 segundos.
2. O tempo de replicação do SyncMirror expirou e as operações continuam no local principal, deixando a réplica remota obsoleta.
3. O site principal é perdido. O resultado é a presença de alterações não replicadas no site principal. Uma mudança pode então ser indesejável por uma série de razões, incluindo o seguinte:
    - Dados críticos podem estar presentes no site principal e esses dados podem eventualmente ser recuperáveis. Um switchover que permitiu que o aplicativo continuasse operando descartaria efetivamente esses dados críticos.
    - Um aplicativo no site que estava usando recursos de armazenamento no site principal no momento da perda do site pode ter dados em cache. Um switchover introduziria uma versão obsoleta dos dados que não corresponde ao cache.
    - Um sistema operacional no site sobrevivente que estava usando recursos de armazenamento no site principal no momento da perda do site pode ter dados em cache. Um switchover introduziria uma versão obsoleta dos dados que não corresponde ao cache. A opção mais segura é configurar o tiebreaker para enviar um alerta se ele detetar falha no local e, em seguida, fazer com que uma pessoa tome uma decisão sobre se deve forçar um switchover. Os aplicativos e/ou sistemas operacionais podem precisar primeiro ser desligados para limpar os dados armazenados em cache. Além disso, as configurações NVFAIL podem ser usadas para adicionar mais proteção e ajudar a simplificar o processo de failover.

Mediador ONTAP com MetroCluster IP

O Mediador ONTAP é usado com MetroCluster IP e outras soluções ONTAP. Ele funciona como um serviço de desempate tradicional, assim como o software de desempate do MetroCluster discutido acima, mas também inclui um recurso crítico: Executar o switchover automatizado sem supervisão.

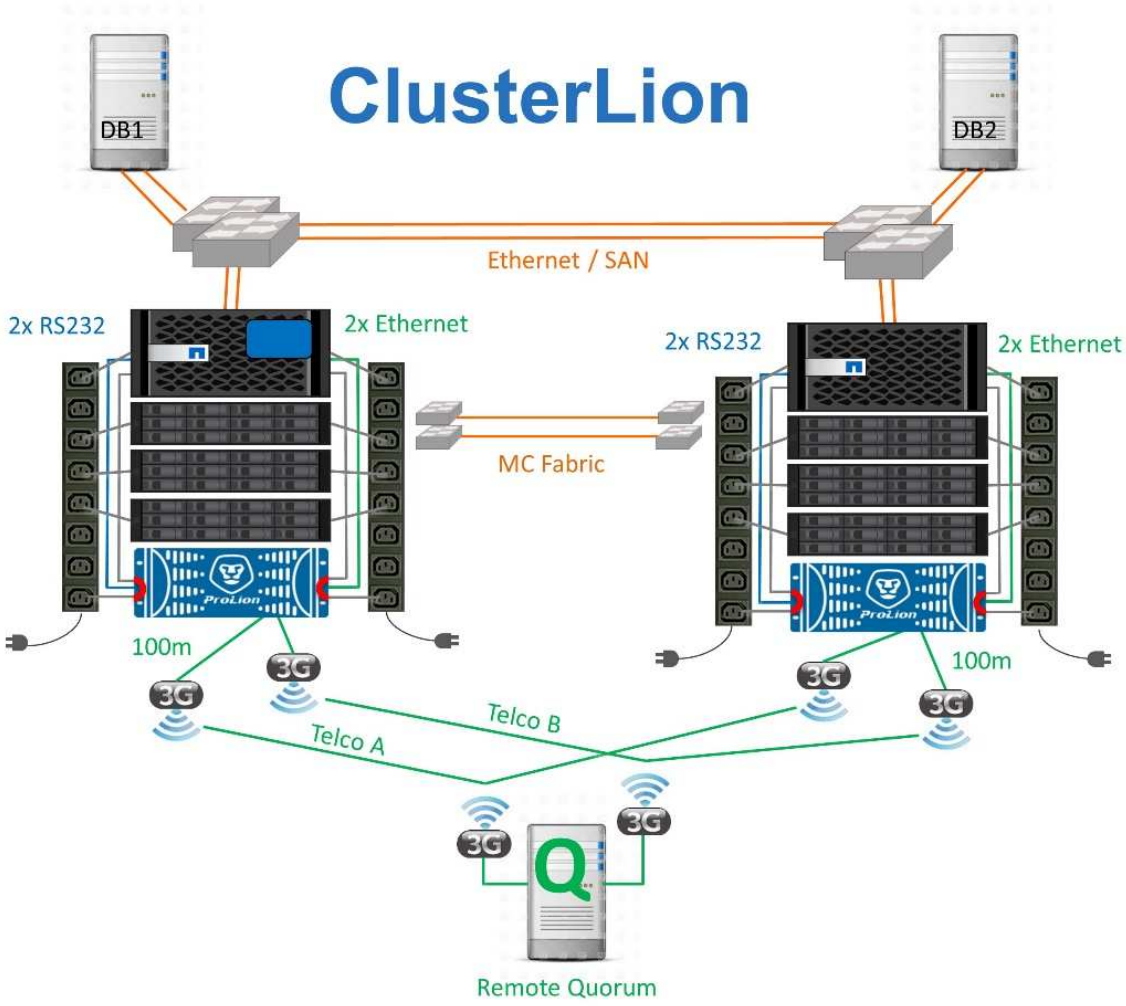
Um MetroCluster conectado à malha tem acesso direto aos dispositivos de storage no local oposto. Isso permite que um controlador MetroCluster monitore a integridade dos outros controladores lendo dados de batimentos cardíacos das unidades. Isso permite que um controlador reconheça a falha de outro controlador e execute um switchover.

Em contraste, a arquitetura IP do MetroCluster roteia todas as I/o exclusivamente através da conexão controlador-controlador; não há acesso direto a dispositivos de armazenamento no local remoto. Isso limita a capacidade de um controlador detetar falhas e executar um switchover. O Mediador ONTAP é, portanto, necessário como um dispositivo de desempate para detetar a perda do local e executar automaticamente um switchover.

Terceiro site virtual com ClusterLion

O ClusterLion é um dispositivo avançado de monitoramento MetroCluster que funciona como um terceiro site virtual. Essa abordagem permite que o MetroCluster seja implantado com segurança em uma configuração de

dois locais com recurso de switchover totalmente automatizado. Além disso, o ClusterLion pode executar um monitor de nível de rede adicional e executar operações pós-switchover. A documentação completa está disponível no ProLion.



- Os dispositivos ClusterLion monitoram a integridade dos controladores com cabos Ethernet e seriais conetados diretamente.
- Os dois aparelhos estão conetados entre si com conexões sem fio redundantes de 3G GHz.
- A alimentação para o controlador ONTAP é direcionada através de relés internos. No caso de uma falha no local, o ClusterLion, que contém um sistema interno de UPS, corta as conexões de energia antes de chamar uma mudança. Este processo garante que nenhuma condição de divisão cerebral ocorra.
- O ClusterLion executa um switchover dentro do tempo limite de 30 segundos do SyncMirror ou não.
- O ClusterLion não executa uma mudança a menos que os estados dos plexes NVRAM e SyncMirror estejam sincronizados.
- Como o ClusterLion só executa um switchover se o MetroCluster estiver totalmente sincronizado, o NVFAIL não é necessário. Essa configuração permite que ambientes que abrangem o local, como um Oracle RAC estendido, permaneçam on-line, mesmo durante um switchover não planejado.
- O suporte inclui MetroCluster conectado à malha e MetroCluster IP

SyncMirror

A base da proteção de dados Oracle com um sistema MetroCluster é o SyncMirror, uma tecnologia de espelhamento síncrono com escalabilidade horizontal e alta performance.

Proteção de dados com o SyncMirror

No nível mais simples, a replicação síncrona significa que qualquer alteração deve ser feita em ambos os lados do storage espelhado antes que seja reconhecida. Por exemplo, se um banco de dados estiver escrevendo um log ou se um convidado VMware estiver sendo corrigido, uma gravação nunca deve ser perdida. Como um nível de protocolo, o sistema de storage não deve reconhecer a gravação até que ela tenha sido comprometida com a Mídia não volátil em ambos os locais. Só então é seguro prosseguir sem o risco de perda de dados.

O uso de uma tecnologia de replicação síncrona é a primeira etapa no projeto e gerenciamento de uma solução de replicação síncrona. A consideração mais importante é entender o que poderia acontecer durante vários cenários de falha planejados e não planejados. Nem todas as soluções de replicação síncrona oferecem os mesmos recursos. Se você precisa de uma solução que forneça um objetivo de ponto de restauração (RPO) zero, o que significa perda de dados zero, é necessário considerar todos os cenários de falha. Em particular, qual é o resultado esperado quando a replicação é impossível devido à perda de conetividade entre sites?

Disponibilidade de dados do SyncMirror

A replicação do MetroCluster é baseada na tecnologia NetApp SyncMirror, projetada para entrar e sair do modo síncrono com eficiência. Essa funcionalidade atende aos requisitos dos clientes que exigem replicação síncrona, mas que também precisam de alta disponibilidade para seus serviços de dados. Por exemplo, se a conetividade a um local remoto for cortada, geralmente é preferível que o sistema de armazenamento continue operando em um estado não replicado.

Muitas soluções de replicação síncrona só são capazes de operar no modo síncrono. Esse tipo de replicação tudo ou nada é às vezes chamado de modo domino. Esses sistemas de storage param de fornecer dados em vez de permitir que cópias locais e remotas dos dados fiquem não sincronizadas. Se a replicação for violada à força, a ressincronização pode ser extremamente demorada e pode deixar um cliente exposto à perda completa de dados durante o tempo em que o espelhamento é restabelecido.

O SyncMirror não só pode alternar facilmente do modo síncrono se o local remoto não estiver acessível, como também pode sincronizar rapidamente para um estado RPO de 0 quando a conetividade é restaurada. A cópia obsoleta dos dados no local remoto também pode ser preservada em um estado utilizável durante a ressincronização, o que garante que cópias locais e remotas dos dados existam em todos os momentos.

Quando o modo domino é necessário, o NetApp oferece SnapMirror Synchronous (SM-S). Opções de nível de aplicativo também existem, como o Oracle DataGuard ou o SQL Server Always On Availability Groups. O espelhamento de disco no nível DO SO pode ser uma opção. Consulte sua equipe de conta do NetApp ou do parceiro para obter informações e opções adicionais.

MetroCluster e NVFAIL

O NVFAIL é um recurso de integridade de dados geral do ONTAP projetado para maximizar a proteção da integridade de dados com bancos de dados.

Esta seção expande a explicação do NVFAIL básico do ONTAP para cobrir tópicos específicos do MetroCluster.

Com o MetroCluster, uma gravação não é reconhecida até que ela tenha sido registrada no NVRAM local e no NVRAM em pelo menos um outro controlador. Essa abordagem garante que uma falha de hardware ou falha de energia não resulte na perda de e/S em trânsito Se o NVRAM local falhar ou a conectividade com outros nós falhar, os dados não serão mais espelhados.

Se o NVRAM local relatar um erro, o nó será encerrado. Esse desligamento resulta em failover para uma controladora de parceiro quando os pares de HA são usados. Com o MetroCluster, o comportamento depende da configuração geral escolhida, mas pode resultar em failover automático para a nota remota. Em qualquer caso, nenhum dado é perdido porque o controlador que está tendo a falha não reconheceu a operação de gravação.

Uma falha de conectividade local a local que bloqueia a replicação do NVRAM para nós remotos é uma situação mais complicada. As gravações não são mais replicadas nos nós remotos, criando uma possibilidade de perda de dados se ocorrer um erro catastrófico em um controlador. Mais importante ainda, tentar fazer failover para um nó diferente durante essas condições resulta em perda de dados.

O fator de controle é se o NVRAM está sincronizado. Se o NVRAM estiver sincronizado, o failover de nó para nó será seguro para prosseguir sem o risco de perda de dados. Em uma configuração do MetroCluster, se o NVRAM e os plexos agregados subjacentes estiverem sincronizados, é seguro prosseguir com o switchover sem o risco de perda de dados.

O ONTAP não permite um failover ou switchover quando os dados estão fora de sincronia, a menos que o failover ou switchover seja forçado. Forçar uma alteração de condições desta forma reconhece que os dados podem ser deixados para trás no controlador original e que a perda de dados é aceitável.

Os bancos de dados são especialmente vulneráveis à corrupção se um failover ou switchover for forçado porque os bancos de dados mantêm caches internos maiores de dados no disco. Se ocorrer um failover forçado ou switchover, as alterações anteriormente confirmadas serão efetivamente descartadas. O conteúdo da matriz de armazenamento salta efetivamente para trás no tempo, e o estado do cache do banco de dados não reflete mais o estado dos dados no disco.

Para proteger aplicativos contra essa situação, o ONTAP permite que volumes sejam configurados para proteção especial contra falha do NVRAM. Quando acionado, esse mecanismo de proteção resulta em um volume entrando em um estado chamado NVFAIL. Esse estado resulta em erros de e/S que causam o desligamento de um aplicativo para que eles não usem dados obsoletos. Os dados não devem ser perdidos porque quaisquer gravações reconhecidas ainda estão presentes no sistema de armazenamento e, com bancos de dados, quaisquer dados de transações confirmadas devem estar presentes nos logs.

As próximas etapas usuais são para que um administrador desligue totalmente os hosts antes de colocar manualmente os LUNs e volumes novamente on-line. Embora essas etapas possam envolver algum trabalho, essa abordagem é a maneira mais segura de garantir a integridade dos dados. Nem todos os dados exigem essa proteção, e é por isso que o comportamento do NVFAIL pode ser configurado volume a volume.

NVFAIL forçado manualmente

A opção mais segura para forçar um switchover com um cluster de aplicativos (incluindo VMware, Oracle RAC e outros) que é distribuído entre locais é especificando `-force-nvfail-all` na linha de comando. Essa opção está disponível como uma medida de emergência para garantir que todos os dados em cache sejam limpos. Se um host estiver usando recursos de armazenamento localizados originalmente no local afetado por desastre, ele receberá erros de e/S ou um (`ESTALE`erro de identificador de arquivo obsoleto). Os bancos de dados Oracle falham e os sistemas de arquivos ficam totalmente offline ou mudam para o modo somente

leitura.

Após a conclusão do switchover, o `in-nvfailed-state` sinalizador precisa ser limpo e os LUNs precisam ser colocados on-line. Depois de concluir esta atividade, a base de dados pode ser reiniciada. Essas tarefas podem ser automatizadas para reduzir o rto.

dr-force-nvfail

Como medida geral de segurança, defina o `dr-force-nvfail` sinalizador em todos os volumes que possam ser acessados de um local remoto durante operações normais, o que significa que são atividades usadas antes do failover. O resultado desta definição é que os volumes remotos selecionados ficam indisponíveis quando entram `in-nvfailed-state` durante um switchover. Após a conclusão do switchover, o `in-nvfailed-state` sinalizador deve ser limpo e os LUNs devem ser colocados on-line. Depois de concluir estas atividades, as aplicações podem ser reiniciadas. Essas tarefas podem ser automatizadas para reduzir o rto.

O resultado é como usar a `-force-nvfail-all` bandeira para switchovers manuais. No entanto, o número de volumes afetados pode ser limitado apenas aos volumes que devem ser protegidos de aplicativos ou sistemas operacionais com caches obsoletos.

Há dois requisitos essenciais para um ambiente que não é usado `dr-force-nvfail` em volumes de aplicações:

- Um switchover forçado não deve ocorrer mais de 30 segundos após a perda do local principal.
- Um switchover não deve ocorrer durante as tarefas de manutenção ou quaisquer outras condições em que os plexos SyncMirror ou a replicação NVRAM estejam fora de sincronia. O primeiro requisito pode ser atendido usando o software tiebreaker configurado para executar um switchover em 30 segundos após uma falha no local. Esse requisito não significa que o switchover deve ser executado dentro de 30 segundos após a deteção de uma falha no local. Isso significa que não é mais seguro forçar uma mudança se tiverem decorrido 30 segundos desde que um local foi confirmado como operacional.

O segundo requisito pode ser parcialmente atendido desativando todos os recursos de switchover automatizado quando a configuração do MetroCluster estiver fora de sincronia. Uma opção melhor é ter uma solução de desempate que possa monitorar a integridade da replicação do NVRAM e dos plexes do SyncMirror. Se o cluster não estiver totalmente sincronizado, o desempate não deverá acionar um switchover.

O software MCTB da NetApp não consegue monitorizar o estado da sincronização, pelo que deve ser desativado quando o MetroCluster não está sincronizado por qualquer motivo. O ClusterLion inclui recursos de monitoramento NVRAM e monitoramento Plex e pode ser configurado para não acionar o switchover, a menos que o sistema MetroCluster seja confirmado como totalmente sincronizado.

Instância única Oracle

Como dito anteriormente, a presença de um sistema MetroCluster não necessariamente adiciona ou altera quaisquer práticas recomendadas para operar um banco de dados. A maioria dos bancos de dados atualmente em execução em sistemas MetroCluster cliente é uma instância única e segue as recomendações na documentação do Oracle On ONTAP.

Failover com um SO pré-configurado

O SyncMirror fornece uma cópia síncrona dos dados no local de recuperação de desastre, mas disponibilizar esses dados requer um sistema operacional e as aplicações associadas. A automação básica pode melhorar significativamente o tempo de failover do ambiente geral. Os produtos Clusterware, como o Veritas Cluster Server (VCS), são frequentemente usados para criar um cluster nos sites e, em muitos casos, o processo de failover pode ser conduzido com scripts simples.

Se os nós primários forem perdidos, o clusterware (ou scripts) é configurado para colocar os bancos de dados on-line no site alternativo. Uma opção é criar servidores de reserva pré-configurados para os recursos NFS ou SAN que compõem o banco de dados. Se o site principal falhar, a alternativa clusterware ou scripted executa uma sequência de ações semelhantes às seguintes:

1. Forçar um switchover do MetroCluster
2. Realizando a descoberta de FC LUNs (somente SAN)
3. Montagem de sistemas de arquivos e/ou montagem de grupos de discos ASM
4. Iniciando o banco de dados

O principal requisito dessa abordagem é um sistema operacional em execução no local remoto. Ele deve ser pré-configurado com binários Oracle, o que também significa que tarefas como patches Oracle devem ser executadas no site primário e em espera. Como alternativa, os binários Oracle podem ser espelhados para o local remoto e montados se um desastre for declarado.

O procedimento de ativação real é simples. Comandos como o reconhecimento LUN requerem apenas alguns comandos por porta FC. A montagem do sistema de arquivos não é mais do que um `mount` comando, e os bancos de dados e ASM podem ser iniciados e parados na CLI com um único comando. Se os volumes e os sistemas de arquivos não estiverem em uso no local de recuperação de desastres antes do switchover, não há requisito de definir `dr-force- nvfail` os volumes.

Failover com um sistema operacional virtualizado

O failover de ambientes de banco de dados pode ser estendido para incluir o próprio sistema operacional. Em teoria, esse failover pode ser feito com LUNs de inicialização, mas na maioria das vezes é feito com um sistema operacional virtualizado. O procedimento é semelhante aos seguintes passos:

1. Forçar um switchover do MetroCluster
2. Montagem dos armazenamentos de dados que hospedam as máquinas virtuais do servidor de banco de dados
3. Iniciar as máquinas virtuais
4. Iniciando bancos de dados manualmente ou configurando as máquinas virtuais para iniciar automaticamente os bancos de dados, por exemplo, um cluster ESX pode abranger sites. Em caso de desastre, as máquinas virtuais podem ser colocadas on-line no local de recuperação de desastres após o switchover. Desde que os armazenamentos de dados que hospedam os servidores de banco de dados virtualizados não estejam em uso no momento do desastre, não há nenhum requisito para definir `dr-force- nvfail` os volumes associados.

Oracle Extended RAC

Muitos clientes otimizam seu rto alongando um cluster do Oracle RAC entre locais, gerando uma configuração totalmente ativo-ativo. O projeto geral se torna mais

complicado porque deve incluir o gerenciamento de quórum do Oracle RAC. Além disso, os dados são acessados de ambos os sites, o que significa que uma mudança forçada pode levar ao uso de uma cópia desatualizada dos dados.

Embora uma cópia dos dados esteja presente em ambos os sites, apenas o controlador que atualmente possui um agregado pode fornecer dados. Portanto, com clusters RAC estendidos, os nós remotos devem executar e/S em uma conexão local a local. O resultado é uma latência de e/S adicionada, mas essa latência geralmente não é um problema. A rede de interconexão RAC também deve ser estendida entre locais, o que significa que uma rede de alta velocidade e baixa latência é necessária de qualquer maneira. Se a latência adicionada causar um problema, o cluster pode ser operado de forma ativo-passivo. As operações com uso intenso de e/S precisariam então ser direcionadas para os nós RAC que são locais para a controladora que possui os agregados. Os nós remotos executam então operações de e/S mais leves ou são usados puramente como servidores de espera quentes.

Se o RAC estendido ativo-ativo for necessário, a sincronização ativa do SnapMirror deve ser considerada no lugar do MetroCluster. A replicação SM-as permite que uma réplica específica dos dados seja preferida. Portanto, um cluster RAC estendido pode ser construído no qual todas as leituras ocorrem localmente. A I/o de leitura nunca cruza sites, o que proporciona a menor latência possível. Todas as atividades de gravação ainda devem transitar a conexão entre locais, mas esse tráfego é inevitável com qualquer solução de espelhamento síncrono.

> Se os LUNs de inicialização, incluindo discos de inicialização virtualizados, forem usados com o Oracle RAC, o `misscount` parâmetro pode precisar ser alterado. Para obter mais informações sobre os parâmetros de tempo limite do RAC, "Oracle RAC com ONTAP"consulte .

Configuração de dois locais

Uma configuração RAC estendida de dois locais pode fornecer serviços de banco de dados ativo-ativo que podem sobreviver a muitos, mas não todos, cenários de desastre sem interrupções.

Ficheiros de votação RAC

A primeira consideração ao implantar o RAC estendido no MetroCluster deve ser o gerenciamento de quórum. O Oracle RAC tem dois mecanismos para gerenciar quórum: O batimento cardíaco do disco e o batimento cardíaco da rede. O heartbeat do disco monitora o acesso ao armazenamento usando os arquivos de votação. Com uma configuração RAC de local único, um único recurso de votação é suficiente, desde que o sistema de armazenamento subjacente ofereça recursos de HA.

Em versões anteriores do Oracle, os arquivos de votação foram colocados em dispositivos de armazenamento físico, mas nas versões atuais do Oracle os arquivos de votação são armazenados em grupos de discos ASM.

> O Oracle RAC é compatível com NFS. Durante o processo de instalação da grade, um conjunto de processos ASM é criado para apresentar o local NFS usado para arquivos de grade como um grupo de discos ASM. O processo é quase transparente para o usuário final e não requer gerenciamento contínuo do ASM após a conclusão da instalação.

O primeiro requisito em uma configuração de dois locais é garantir que cada local possa sempre acessar mais da metade dos arquivos de votação de forma que garanta um processo de recuperação de desastres sem interrupções. Essa tarefa era simples antes que os arquivos de votação fossem armazenados em grupos de discos ASM, mas hoje os administradores precisam entender os princípios básicos da redundância ASM.

Os grupos de discos ASM têm três opções para redundância `external`, `normal` e `high`. Em outras palavras, sem espelhamento, espelhado e espelhado de 3 vias. Uma opção mais recente chamada `Flex`

também está disponível, mas raramente usada. O nível de redundância e o posicionamento dos dispositivos redundantes controlam o que acontece em cenários de falha. Por exemplo:

- Colocar os arquivos de votação em um `diskgroup` recurso com `external` redundância garante despejo de um site se a conetividade entre sites for perdida.
- Colocar os arquivos de votação em um `diskgroup` com `normal` redundância com apenas um disco ASM por site garante despejo de nó em ambos os sites se a conetividade entre sites for perdida porque nenhum dos sites teria quórum de maioria.
- Colocar os arquivos de votação em um `diskgroup` `high` com redundância com dois discos em um local e um único disco no outro local permite operações ativas-ativas quando ambos os sites estão operacionais e mutuamente acessíveis. No entanto, se o site de disco único for isolado da rede, então esse site é despejado.

Batimento cardíaco da rede RAC

O batimento cardíaco da rede do Oracle RAC monitora a acessibilidade dos nós na interconexão de cluster. Para permanecer no cluster, um nó deve ser capaz de entrar em Contato com mais da metade dos outros nós. Em uma arquitetura de dois locais, esse requisito cria as seguintes opções para a contagem de nós RAC:

- O posicionamento de um número igual de nós por local resulta em despejo em um local caso a conetividade de rede seja perdida.
- O posicionamento de nós N em um local e nós N-1 no outro local garante que a perda de conetividade entre locais resulta no local com o maior número de nós restantes no quórum de rede e o local com menos nós despejando.

Antes do Oracle 12cR2, não era viável controlar qual lado experimentaria um despejo durante a perda do local. Quando cada local tem um número igual de nós, o despejo é controlado pelo nó principal, que em geral é o primeiro nó RAC a inicializar.

O Oracle 12cR2 introduz a capacidade de ponderação de nós. Essa capacidade dá a um administrador mais controle sobre como a Oracle resolve condições de split-brain. Como um exemplo simples, o comando a seguir define a preferência de um nó específico em um RAC:

```
[root@host-a ~]# /grid/bin/crsctl set server css_critical yes
CRS-4416: Server attribute 'CSS_CRITICAL' successfully changed. Restart
Oracle High Availability Services for new value to take effect.
```

Após reiniciar o Oracle High-Availability Services, a configuração será a seguinte:

```
[root@host-a lib]# /grid/bin/crsctl status server -f | egrep
'^NAME|CSS_CRITICAL='
NAME=host-a
CSS_CRITICAL=yes
NAME=host-b
CSS_CRITICAL=no
```

O nó `host-a` agora é designado como o servidor crítico. Se os dois nós RAC estiverem isolados, `host-a` sobrevive e `host-b` é despejado.

Para obter detalhes completos, consulte o white paper da Oracle "Visão geral técnica do Oracle Clusterware 12c versão 2. "

Para versões do Oracle RAC anteriores ao 12cR2, o nó principal pode ser identificado verificando os logs do CRS da seguinte forma:

```
[root@host-a ~]# /grid/bin/crsctl status server -f | egrep
'^NAME|CSS_CRITICAL='
NAME=host-a
CSS_CRITICAL=yes
NAME=host-b
CSS_CRITICAL=no
 [root@host-a ~]# grep -i 'master node' /grid/diag/crs/host-
a/crs/trace/crsd.trc
2017-05-04 04:46:12.261525 :   CRSSE:2130671360: {1:16377:2} Master Change
Event; New Master Node ID:1 This Node's ID:1
2017-05-04 05:01:24.979716 :   CRSSE:2031576832: {1:13237:2} Master Change
Event; New Master Node ID:2 This Node's ID:1
2017-05-04 05:11:22.995707 :   CRSSE:2031576832: {1:13237:221} Master
Change Event; New Master Node ID:1 This Node's ID:1
2017-05-04 05:28:25.797860 :   CRSSE:3336529664: {1:8557:2} Master Change
Event; New Master Node ID:2 This Node's ID:1
```

Este log indica que o nó principal é `2` e o nó `host-a` tem uma ID de `1`. Esse fato significa que `host-a` não é o nó principal. A identidade do nó principal pode ser confirmada com o comando `olsnodes -n`.

```
[root@host-a ~]# /grid/bin/olsnodes -n
host-a  1
host-b  2
```

O nó com uma ID de `2` é `host-b`, que é o nó principal. Em uma configuração com números iguais de nós em cada site, o site com `host-b` é o site que sobrevive se os dois conjuntos perderem a conetividade de rede por qualquer motivo.

É possível que a entrada de log que identifica o nó mestre possa ficar fora do sistema. Nesta situação, os carimbos de data/hora dos backups do Oracle Cluster Registry (OCR) podem ser usados.

```
[root@host-a ~]#  /grid/bin/ocrconfig -showbackup
host-b     2017/05/05 05:39:53     /grid/cdata/host-cluster/backup00.ocr
0
host-b     2017/05/05 01:39:53     /grid/cdata/host-cluster/backup01.ocr
0
host-b     2017/05/04 21:39:52     /grid/cdata/host-cluster/backup02.ocr
0
host-a     2017/05/04 02:05:36     /grid/cdata/host-cluster/day.ocr     0
host-a     2017/04/22 02:05:17     /grid/cdata/host-cluster/week.ocr     0
```

Este exemplo mostra que o nó principal é `host-b`. Ele também indica uma mudança no nó mestre de `host-a` para `host-b` algum lugar entre 2:05 e 21:39 em 4 de maio. Este método de identificação do nó principal só é seguro se os logs do CRS também tiverem sido verificados porque é possível que o nó principal tenha sido alterado desde o backup OCR anterior. Se essa alteração tiver ocorrido, ela deverá estar visível nos logs do OCR.

A maioria dos clientes escolhe um único grupo de discos de votação que atende todo o ambiente e um número igual de nós RAC em cada local. O grupo de discos deve ser colocado no site que contém o banco de dados. O resultado é que a perda de conetividade resulta em despejo no local remoto. O site remoto não teria mais quórum, nem teria acesso aos arquivos do banco de dados, mas o site local continua sendo executado como de costume. Quando a conetividade é restaurada, a instância remota pode ser colocada online novamente.

Em caso de desastre, é necessário um switchover para colocar os arquivos do banco de dados e o grupo de discos de votação on-line no local sobrevivente. Se o desastre permitir que o AUSO acione o switchover, o NVFAIL não será acionado porque o cluster é conhecido por estar em sincronia e os recursos de storage ficam online normalmente. AUSO é uma operação muito rápida e deve ser concluída antes que o `disktimeout` período expire.

Como existem apenas dois locais, não é possível usar qualquer tipo de software de quebra de informações externo automatizado, o que significa que o switchover forçado deve ser uma operação manual.

Configurações de três locais

Um cluster RAC estendido é muito mais fácil de arquitetar com três locais. Os dois sites que hospedam cada metade do sistema MetroCluster também dão suporte aos workloads de banco de dados, enquanto o terceiro local serve como desempate para o banco de dados e para o sistema MetroCluster. A configuração do Oracle tiebreaker pode ser tão simples quanto colocar um membro do grupo de discos ASM usado para votar em um site 3rd e também pode incluir uma instância operacional no site 3rd para garantir que haja um número ímpar de nós no cluster RAC.

Consulte a documentação da Oracle sobre "grupo de falha de quórum" para obter informações importantes sobre o uso do NFS em uma configuração RAC estendida. Em resumo, as opções de montagem NFS podem precisar ser modificadas para incluir a opção de software para garantir que a perda de conetividade com os recursos de quórum de hospedagem de sites 3rd não pendure os servidores Oracle primários ou os processos Oracle RAC.

Informações sobre direitos autorais

Informações sobre marcas comerciais