Arquitetura de alta disponibilidade

ONTAP Select

NetApp
January 31, 2025

Índice

Arquitetura de alta disponibilidade

Configurações de alta disponibilidade

Descubra as opções de alta disponibilidade para selecionar a melhor configuração de HA para o seu ambiente.

Embora os clientes estejam começando a migrar workloads de aplicações de dispositivos de storage de classe empresarial para soluções baseadas em software executadas em hardware comum, as expectativas e necessidades relacionadas à resiliência e à tolerância de falhas não mudaram. Uma solução de HA que fornece um objetivo de ponto de restauração zero (RPO) protege o cliente da perda de dados devido a uma falha de qualquer componente do stack de infraestrutura.

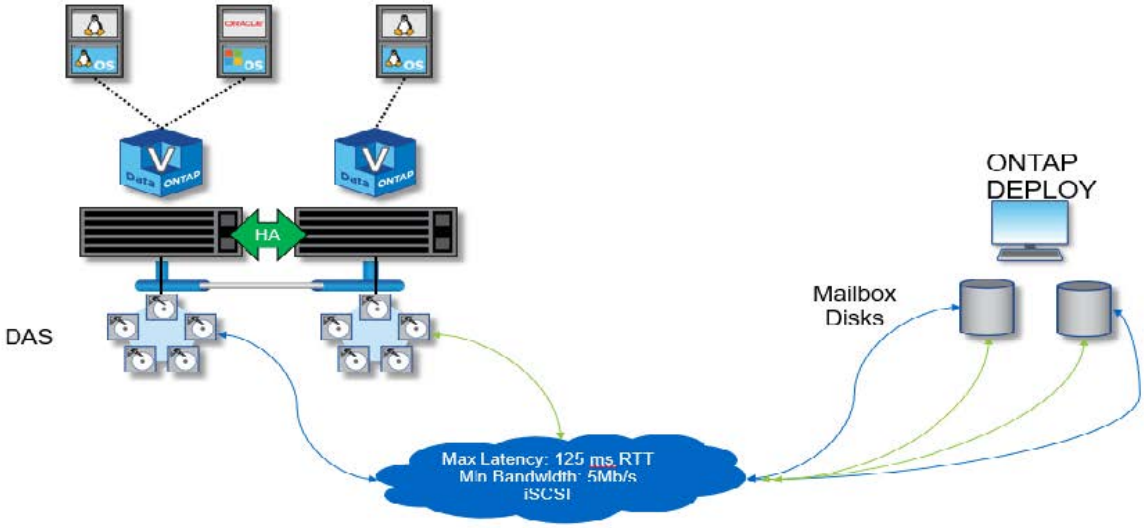
Uma grande parte do mercado de SDS é baseada na noção de storage sem compartilhamento, com a replicação de software fornecendo resiliência de dados ao armazenar várias cópias de dados de usuários em diferentes silos de storage. O ONTAP Select se baseia nessa premissa usando os recursos de replicação síncrona (RAID SyncMirror) fornecidos pelo ONTAP para armazenar uma cópia extra dos dados do usuário no cluster. Isso ocorre no contexto de um par de HA. Cada par de HA armazena duas cópias de dados de usuário: Uma no storage fornecido pelo nó local e outra no storage fornecido pelo parceiro de HA. Em um cluster do ONTAP Select, a replicação síncrona e de HA são Unidas e o recurso dos dois não pode ser desacoplado ou usado de forma independente. Como resultado, a funcionalidade de replicação síncrona só está disponível na oferta multinode.

> Em um cluster ONTAP Select, a funcionalidade de replicação síncrona é uma função da implementação de HA, e não um substituto para os mecanismos de replicação assíncrona SnapMirror ou SnapVault. A replicação síncrona não pode ser usada independentemente do HA.

Há dois modelos de implantação do ONTAP Select HA: Os clusters com vários nós (quatro, seis ou oito nós) e os clusters de dois nós. A caraterística principal de um cluster de ONTAP Select de dois nós é o uso de um serviço de mediador externo para resolver cenários de split-brain. A VM ONTAP Deploy serve como mediador padrão para todos os pares de HA de dois nós que ela configura.

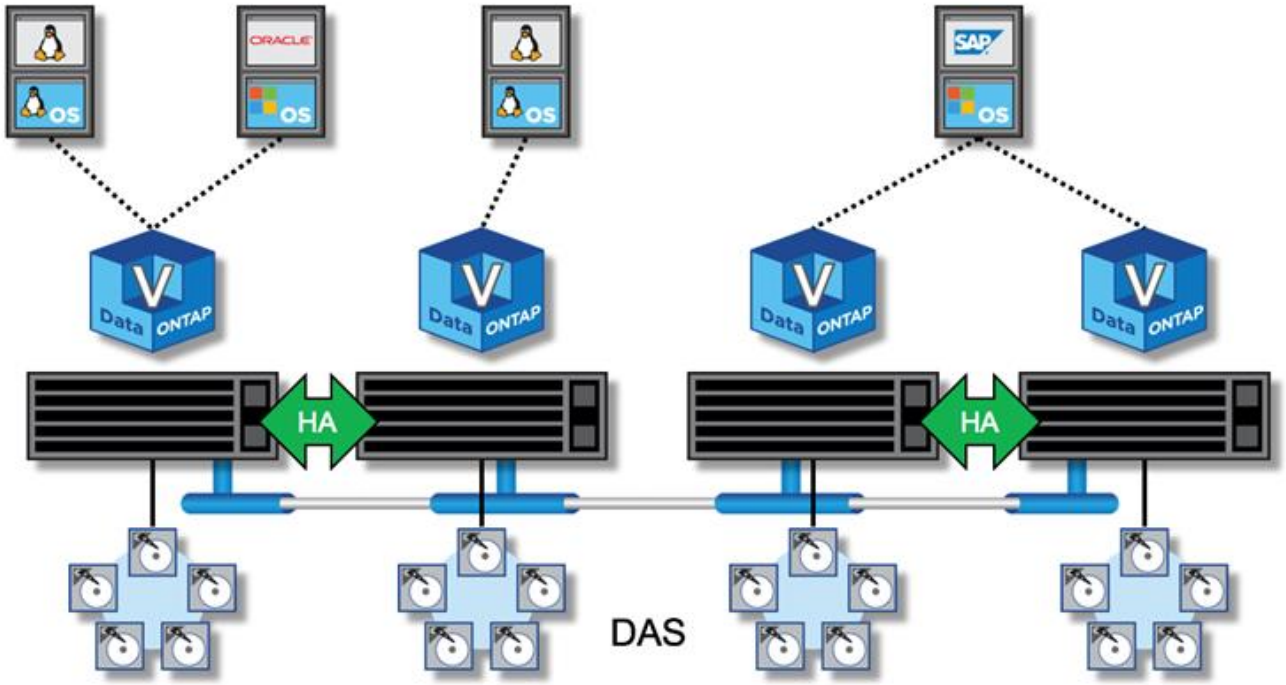
As duas arquiteturas são representadas nas figuras a seguir.

- Cluster ONTAP Select de dois nós com mediador remoto e usando armazenamento conetado local*



O cluster de dois nós do ONTAP Select é composto por um par de HA e um mediador. No par de HA, os agregados de dados em cada nó de cluster são espelhados de forma síncrona e, no caso de um failover, não há perda de dados.

- Cluster ONTAP Select de quatro nós usando armazenamento conetado local*



- O cluster de quatro nós ONTAP Select é composto por dois pares de HA. Os clusters de seis nós e oito nós são compostos por três e quatro pares de HA, respetivamente. Em cada par de HA, os agregados de dados em cada nó de cluster são espelhados de forma síncrona e, em caso de failover, não há perda de dados.
- Apenas uma instância do ONTAP Select pode estar presente em um servidor físico ao usar o armazenamento DAS. O ONTAP Select requer acesso não compartilhado à controladora RAID local do sistema e foi projetado para gerenciar os discos conetados localmente, o que seria impossível sem conetividade física ao storage.

Ha de dois nós versus HA de vários nós

Diferentemente dos arrays FAS, os nós ONTAP Select em um par de HA se comunicam exclusivamente pela rede IP. Isso significa que a rede IP é um único ponto de falha (SPOF), e proteger contra partições de rede e cenários de split-brain torna-se um aspeto importante do projeto. O cluster com vários nós pode sustentar falhas de nó único porque o quorum do cluster pode ser estabelecido pelos três ou mais nós sobreviventes. O cluster de dois nós conta com o serviço de mediador hospedado pela VM ONTAP Deploy para obter o mesmo resultado.

O tráfego de rede Heartbeat entre os nós do ONTAP Select e o serviço de mediador ONTAP Deploy é mínimo e resiliente para que a VM ONTAP Deploy possa ser hospedada em um data center diferente do cluster de dois nós do ONTAP Select.

A VM do ONTAP Deploy se torna parte integrante de um cluster de dois nós quando atua como mediador desse cluster. Se o serviço de mediador não estiver disponível, o cluster de dois nós continuará fornecendo dados, mas os recursos de failover de storage do cluster ONTAP Select serão desativados. Portanto, o serviço de mediador ONTAP Deploy deve manter a comunicação constante com cada nó ONTAP Select no par de HA. Uma largura de banda mínima de 5Mbps Gbps e uma latência máxima de tempo de ida e volta (RTT) de 125ms ms são necessários para permitir o funcionamento adequado do quorum do cluster.

Se a VM de implantação do ONTAP atuando como mediador estiver temporariamente ou potencialmente indisponível permanentemente, uma VM secundária de implantação do ONTAP poderá ser usada para restaurar o quórum de cluster de dois nós. Isso resulta em uma configuração na qual a nova VM de implantação do ONTAP não consegue gerenciar os nós do ONTAP Select, mas participa com êxito do algoritmo de quorum do cluster. A comunicação entre os nós do ONTAP Select e a VM de implantação do ONTAP é feita usando o protocolo iSCSI em IPv4. O endereço IP de gerenciamento do nó ONTAP Select é o iniciador e o endereço IP da VM de implantação do ONTAP é o destino. Portanto, não é possível suportar endereços IPv6 para os endereços IP de gerenciamento de nós ao criar um cluster de dois nós. Os discos da caixa de correio hospedada do ONTAP Deploy são criados automaticamente e mascarados para os endereços IP de gerenciamento de nós do ONTAP Select apropriados no momento da criação do cluster de dois nós. Toda a configuração é executada automaticamente durante a configuração e nenhuma ação administrativa adicional é necessária. A instância do ONTAP Deploy que cria o cluster é o mediador padrão desse cluster.

Uma ação administrativa é necessária se o local original do mediador tiver de ser alterado. É possível recuperar um quórum de cluster mesmo que a VM de implantação original do ONTAP seja perdida. No entanto, a NetApp recomenda que você faça backup do banco de dados ONTAP Deploy depois que cada cluster de dois nós for instanciado.

Ha de dois nós versus HA estendida de dois nós (MetroCluster SDS)

É possível esticar um cluster de HA ativo/ativo de dois nós em distâncias maiores e potencialmente colocar cada nó em um data center diferente. A única distinção entre um cluster de dois nós e um cluster estendido de dois nós (também conhecido como MetroCluster SDS) é a distância de conetividade de rede entre nós.

O cluster de dois nós é definido como um cluster para o qual ambos os nós estão localizados no mesmo data center a uma distância de 300m km. Em geral, ambos os nós têm uplinks para o mesmo switch de rede ou conjunto de switches de rede ISL (Interswitch link).

O MetroCluster SDS de dois nós é definido como um cluster para o qual os nós são separados fisicamente (salas diferentes, edifícios diferentes e data centers diferentes) em mais de 300m. Além disso, as conexões uplink de cada nó são conetadas a switches de rede separados. O SDS do MetroCluster não requer hardware dedicado. No entanto, o ambiente deve aderir aos requisitos de latência (um máximo de 5ms para RTT e 5ms

para jitter, para um total de 10ms) e distância física (um máximo de 10km).

O MetroCluster SDS é um recurso premium e requer uma licença Premium ou uma licença Premium XL. A licença Premium suporta a criação de VMs pequenas e médias, bem como suportes HDD e SSD. A licença Premium XL também dá suporte à criação de unidades NVMe.

> O MetroCluster SDS é compatível com storage anexado local (DAS) e storage compartilhado (vNAS). Observe que as configurações do vNAS geralmente têm uma latência inata maior devido à rede entre a VM do ONTAP Select e o armazenamento compartilhado. As configurações do MetroCluster SDS devem fornecer um máximo de 10ms ms de latência entre os nós, incluindo a latência de storage compartilhado. Em outras palavras, apenas medir a latência entre as VMs Select não é adequada, pois a latência de armazenamento compartilhado não é insignificante para essas configurações.

RSM DE HA e agregados espelhados

Evite a perda de dados usando RAID SyncMirror (RSM), agregados espelhados e caminho de gravação.

Replicação síncrona

O modelo ONTAP HA foi desenvolvido com base no conceito de parceiros de HA. O ONTAP Select estende essa arquitetura para o mundo dos servidores comuns não compartilhados usando o recurso RAID SyncMirror (RSM) presente no ONTAP para replicar blocos de dados entre nós do cluster, fornecendo duas cópias de dados de usuários espalhados por um par de HA.

Um cluster de dois nós com um mediador pode abranger dois data centers. Para obter mais informações, consulte a "Práticas recomendadas de HA (MetroCluster SDS) com dois nós esticados"seção .

Agregados espelhados

Um cluster do ONTAP Select é composto por dois a oito nós. Cada par de HA contém duas cópias de dados de usuário, espelhadas sincronamente entre nós em uma rede IP. Esse espelhamento é transparente para o usuário e é uma propriedade do agregado de dados, configurado automaticamente durante o processo de criação de agregados de dados.

Todos os agregados em um cluster ONTAP Select devem ser espelhados para disponibilidade de dados em caso de failover de nó e para evitar um SPOF em caso de falha de hardware. Os agregados em um cluster ONTAP Select são criados a partir de discos virtuais fornecidos de cada nó no par de HA e usam os seguintes discos:

- Um conjunto local de discos (fornecido pelo nó ONTAP Select atual)
- Um conjunto espelhado de discos (fornecido pelo parceiro de HA do nó atual)

> Os discos locais e espelhados usados para construir um agregado espelhado devem ter o mesmo tamanho. Estes agregados são referidos como Plex 0 e Plex 1 (para indicar os pares de espelhos locais e remotos, respetivamente). Os números de Plex reais podem ser diferentes em sua instalação.

Essa abordagem é fundamentalmente diferente da maneira como os clusters ONTAP padrão funcionam. Isso se aplica a todos os discos raiz e de dados dentro do cluster ONTAP Select. O agregado contém cópias de dados locais e espelhadas. Portanto, um agregado que contém N discos virtuais oferece o valor de

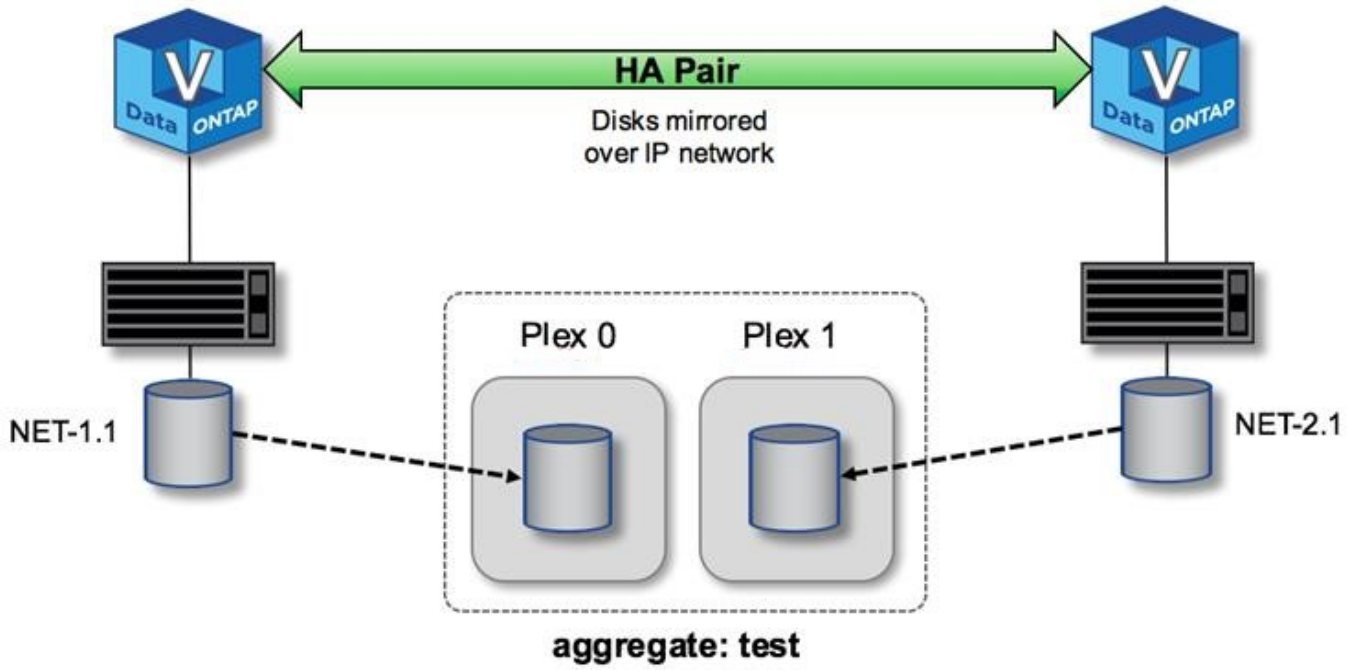
armazenamento exclusivo de discos N/2, porque a segunda cópia de dados reside em seus próprios discos exclusivos.

A figura a seguir mostra um par de HA em um cluster ONTAP Select de quatro nós. Nesse cluster, há um único agregado (teste) que usa o storage de ambos os parceiros de HA. Esse agregado de dados é composto por dois conjuntos de discos virtuais: Um conjunto local, contribuído pelo nó de cluster proprietário do ONTAP Select (Plex 0) e um conjunto remoto, contribuído pelo parceiro de failover (Plex 1).

Plex 0 é o bucket que contém todos os discos locais. Plex 1 é o bucket que contém discos espelhados ou discos responsáveis por armazenar uma segunda cópia replicada dos dados do usuário. O nó que possui o agregado contribui com discos para o Plex 0, e o parceiro de HA desse nó contribui com discos para o Plex 1.

Na figura a seguir, há um agregado espelhado com dois discos. O conteúdo desse agregado é espelhado em nossos dois nós de cluster, com O disco local NET-1,1 colocado no bucket Plex 0 e o disco remoto NET-2,1 colocado no bucket Plex 1. Neste exemplo, o teste agregado é propriedade do nó de cluster à esquerda e usa o disco local NET-1,1 e o disco espelhado do parceiro HA NET-2,1.

Agregado espelhado ONTAP Select



Quando um cluster ONTAP Select é implantado, todos os discos virtuais presentes no sistema são atribuídos automaticamente ao Plex correto, não exigindo nenhuma etapa adicional do usuário em relação à atribuição de disco. Isso impede a atribuição acidental de discos a um Plex incorreto e fornece a configuração ideal do disco espelhado.

Escrever caminho

O espelhamento síncrono de blocos de dados entre os nós do cluster e o requisito para nenhuma perda de dados com uma falha do sistema têm um impacto significativo no caminho que uma gravação recebida leva à medida que se propaga por um cluster ONTAP Select. Este processo consiste em duas etapas:

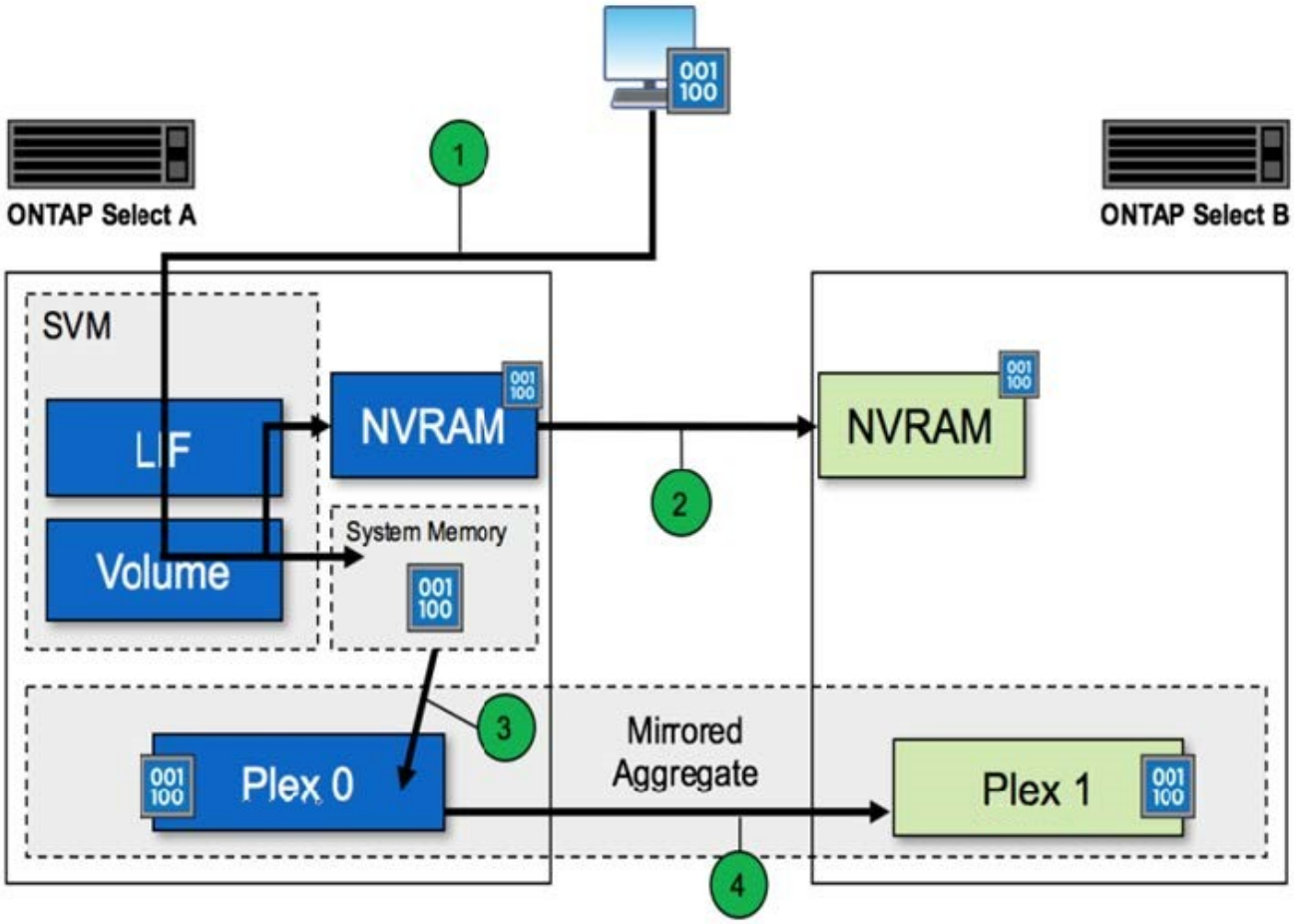
- Confirmação
- Destruição

As gravações em um volume de destino ocorrem em um LIF de dados e são comprometidas com a partição NVRAM virtualizada, presente em um disco do sistema do nó ONTAP Select, antes de serem reconhecidas de volta ao cliente. Em uma configuração de HA, ocorre uma etapa adicional porque essas gravações do NVRAM são espelhadas imediatamente no parceiro de HA do proprietário do volume de destino antes de serem confirmadas. Esse processo garante a consistência do sistema de arquivos no nó do parceiro de HA, se houver uma falha de hardware no nó original.

Depois que a gravação foi comprometida com o NVRAM, o ONTAP move periodicamente o conteúdo desta partição para o disco virtual apropriado, um processo conhecido como destaging. Esse processo só acontece uma vez, no nó do cluster que possui o volume de destino e não acontece no parceiro de HA.

A figura a seguir mostra o caminho de gravação de uma solicitação de gravação recebida em um nó ONTAP Select.

Fluxo de trabalho de caminho de escrita ONTAP Select



A confirmação de gravação recebida inclui as seguintes etapas:

- As gravações entram no sistema através de uma interface lógica de propriedade do nó A. do ONTAP Select
- As gravações são comprometidas com a NVRAM do nó A e espelhadas com o nó B..
- Depois que a solicitação de e/S estiver presente em ambos os nós de HA, a solicitação será então reconhecida de volta ao cliente.

O desarranjo do ONTAP Select do NVRAM para o agregado de dados (ONTAP CP) inclui as seguintes etapas:

- As gravações são destagidas de NVRAM virtual para agregado de dados virtual.
- O motor do espelho replica de forma síncrona os blocos em ambos os plexos.

HA detalhes adicionais

O coração do DISCO HA, a caixa de correio HA, o coração do HA, o failover de HA e o Giveback trabalham para aprimorar a proteção de dados.

Coração do disco batendo

Embora a arquitetura do ONTAP Select HA aproveite muitos dos caminhos de código usados pelos arrays FAS tradicionais, algumas exceções existem. Uma dessas exceções está na implementação do heartbeat baseado em disco, um método de comunicação não baseado em rede usado por nós de cluster para evitar que o isolamento da rede cause comportamento de split-brain. Um cenário de split-brain é o resultado do particionamento de cluster, normalmente causado por falhas de rede, em que cada lado acredita que o outro está inativo e tenta assumir recursos de cluster.

As implementações de HA de classe empresarial devem lidar com esse tipo de cenário de forma graciosa. O ONTAP faz isso por meio de um método personalizado baseado em disco de batimentos cardíacos. Este é o trabalho da caixa de correio de HA, um local no storage físico usado pelos nós de cluster para passar mensagens de batimento cardíaco. Isso ajuda o cluster a determinar a conetividade e, portanto, definir quorum no caso de um failover.

Nos arrays FAS, que usam uma arquitetura de HA de storage compartilhado, o ONTAP resolve problemas de divisão das seguintes maneiras:

- Reservas persistentes de SCSI
- Metadados de HA persistentes
- ESTADO HA enviado por interconexão HA

No entanto, na arquitetura sem compartilhamento de um cluster do ONTAP Select, um nó só consegue ver seu próprio storage local e não o do parceiro de HA. Portanto, quando o particionamento de rede isola cada lado de um par de HA, os métodos anteriores de determinar o quórum de cluster e o comportamento de failover não estão disponíveis.

Embora o método existente de deteção e evitação de split-brain não possa ser usado, um método de mediação ainda é necessário, aquele que se encaixa dentro das restrições de um ambiente de nada compartilhado. O ONTAP Select estende ainda mais a infraestrutura de caixa de correio existente, permitindo que ele atue como um método de mediação em caso de particionamento de rede. Como o armazenamento compartilhado não está disponível, a mediação é realizada por meio do acesso aos discos da caixa de correio através do nas. Esses discos são espalhados pelo cluster, incluindo o mediador em um cluster de dois nós, usando o protocolo iSCSI. Portanto, as decisões de failover inteligentes podem ser tomadas por um nó de cluster com base no acesso a esses discos. Se um nó puder acessar os discos da caixa de correio de outros nós fora de seu parceiro de HA, provavelmente estará ativo e íntegro.

A arquitetura da caixa de correio e o método de heartbeat baseado em disco de resolução de quórum de cluster e problemas de split-brain são os motivos pelos quais a variante multinode do ONTAP Select requer quatro nós separados ou um mediador para um cluster de dois nós.

Postagem da caixa postal HA

A arquitetura da caixa de correio do HA usa um modelo de postagem de mensagens. Em intervalos repetidos, os nós do cluster enviam mensagens para todos os outros discos da caixa de correio no cluster, incluindo o mediador, informando que o nó está ativo e em execução. Em um cluster saudável a qualquer momento, um único disco de caixa de correio em um nó de cluster tem mensagens postadas de todos os outros nós de cluster.

Anexado a cada nó de cluster Select é um disco virtual que é usado especificamente para acesso compartilhado à caixa de correio. Esse disco é chamado de disco de caixa de correio mediador, porque sua principal função é agir como um método de mediação de cluster em caso de falhas de nó ou particionamento de rede. Este disco de caixa de correio contém partições para cada nó de cluster e é montado numa rede iSCSI por outros nós de cluster Select. Periodicamente, esses nós postam status de integridade para a partição apropriada do disco da caixa de correio. O uso de discos de caixa de correio acessíveis à rede espalhados por todo o cluster permite inferir a integridade do nó por meio de uma matriz de acessibilidade. Por exemplo, os nós de cluster A e B podem postar na caixa de correio do nó de cluster D, mas não na caixa de correio do nó C. além disso, o nó de cluster D não pode postar na caixa de correio do nó C, portanto é provável que o nó C esteja inativo ou isolado na rede e deva ser assumido.

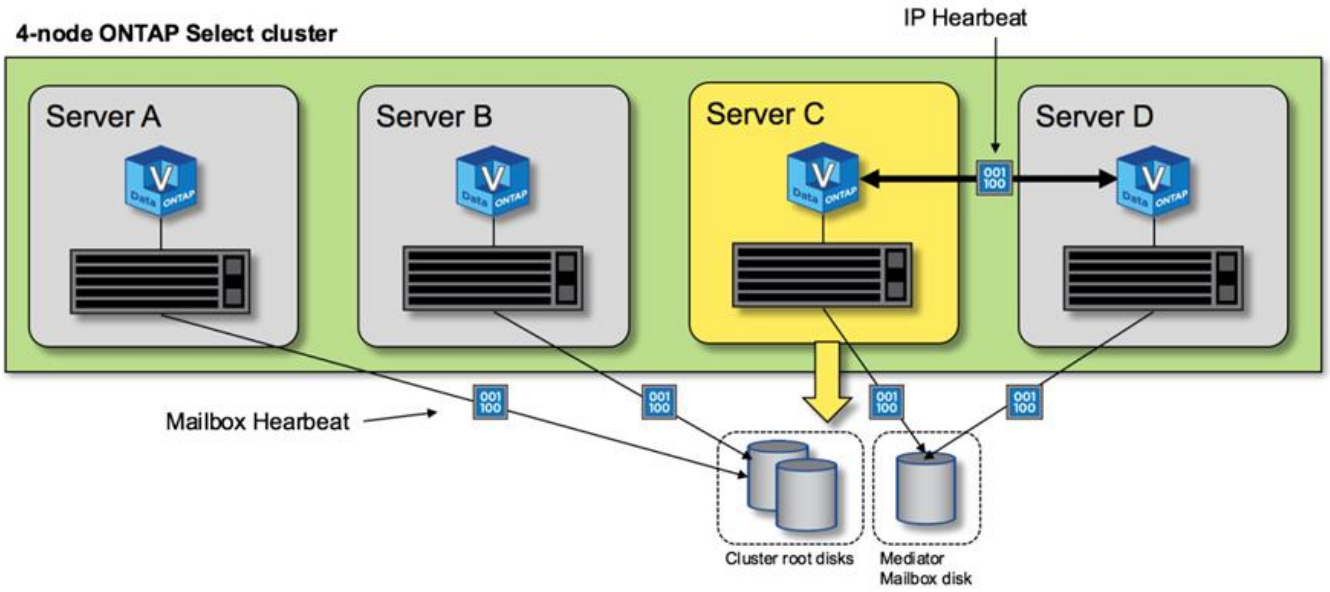
HA coração batendo

Assim como nas plataformas NetApp FAS, o ONTAP Select envia periodicamente mensagens de heartbeat de HA pela interconexão de HA. Dentro do cluster ONTAP Select, isso é realizado por meio de uma conexão de rede TCP/IP que existe entre parceiros de HA. Além disso, as mensagens de heartbeat baseadas em disco são passadas para todos os discos da caixa de correio de HA, incluindo os discos da caixa de correio do mediador. Essas mensagens são passadas a cada poucos segundos e lidas periodicamente. A frequência com que eles são enviados e recebidos permite que o cluster ONTAP Select detete eventos de falha de HA em aproximadamente 15 segundos, a mesma janela disponível nas plataformas FAS. Quando as mensagens de heartbeat não estão mais sendo lidas, um evento de failover é acionado.

A figura a seguir mostra o processo de envio e recebimento de mensagens de heartbeat sobre os discos de interconexão e mediador de HA na perspetiva de um único nó de cluster ONTAP Select, nó C.

Os batimentos cardíacos da rede são enviados pela interconexão de HA para o parceiro de HA, nó D, enquanto os batimentos cardíacos do disco usam discos de caixa postal em todos os nós de cluster, A, B, C e D.

HA heartbearing em um cluster de quatro nós: Estado estável



Failover de HA e giveback

Durante uma operação de failover, o nó sobrevivente assume as responsabilidades de fornecimento de dados para o nó de mesmo nível usando a cópia local dos dados do parceiro de HA. A e/S do cliente pode continuar ininterrupta, mas as alterações a esses dados devem ser replicadas antes que a giveback possa ocorrer. Observe que o ONTAP Select não oferece suporte a um giveback forçado porque isso faz com que as alterações armazenadas no nó sobrevivente sejam perdidas.

A operação de retorno de sincronização é acionada automaticamente quando o nó reinicializado rejura o cluster. O tempo necessário para a sincronização de volta depende de vários fatores. Esses fatores incluem o número de alterações que devem ser replicadas, a latência da rede entre os nós e a velocidade dos subsistemas de disco em cada nó. É possível que o tempo necessário para a sincronização de volta exceda a janela de retorno automático de 10 minutos. Neste caso, é necessário um manual de giveback após a sincronização de volta. O progresso da sincronização de volta pode ser monitorado usando o seguinte comando:

```
storage aggregate status -r -aggregate <aggregate name>
```

Informações sobre direitos autorais

Informações sobre marcas comerciais