Banco de dados Oracle

Enterprise applications

NetApp
February 11, 2026

Índice

Banco de dados Oracle . . . 1
- Bancos de dados Oracle no ONTAP . . . 1
- Configuração do ONTAP em sistemas AFF/ FAS . . . 1
  - RAID . . . 1
  - Gerenciamento de capacidade . . . 2
  - Máquinas virtuais de armazenamento . . . 3
  - Gerenciamento de desempenho com QoS ONTAP . . . 3
  - Eficiência . . . 5
  - Thin Provisioning . . . 9
  - Failover/switchover da ONTAP . . . 11
- Configuração do ONTAP em sistemas ASA r2 . . . 13
  - RAID . . . 13
  - Gerenciamento de capacidade . . . 14
  - Máquinas virtuais de armazenamento . . . 14
  - Gerenciamento de desempenho com ONTAP QoS em sistemas ASA r2 . . . 15
  - Eficiência . . . 16
  - Thin Provisioning . . . 19
  - Failover do ONTAP . . . 20
- Configuração de banco de dados com sistemas AFF/ FAS . . . 22
  - Tamanhos de blocos . . . 22
  - db_file_multiblock_read_count . . . 23
  - sistema de arquivos_options . . . 23
  - Tempos limite do RAC . . . 24
- Configuração de banco de dados com sistemas ASA r2 . . . 26
  - Tamanhos de blocos . . . 26
  - db_file_multiblock_read_count . . . 27
  - sistema de arquivos_options . . . 27
  - Tempos limite do RAC . . . 29
- Configuração de host com sistemas AFF/ FAS . . . 30
  - AIX . . . 30
  - HP-UX . . . 32
  - Linux . . . 34
  - ASMLib/AFD (controlador de filtro ASM) . . . 38
  - Microsoft Windows . . . 40
  - Solaris . . . 40
- Configuração de host com sistemas ASA r2 . . . 46
  - AIX . . . 46
  - HP-UX . . . 47
  - Linux . . . 48
  - ASMLib/AFD (controlador de filtro ASM) . . . 50
  - Microsoft Windows . . . 52
  - Solaris . . . 52
- Configuração de rede em sistemas AFF/ FAS . . . 56

- Interfaces lógicas . . . 56
- Configuração TCP/IP e ethernet . . . 61
- Configuração de FC SAN . . . 62
- Rede de conexão direta . . . 63

Configuração de rede em sistemas ASA r2 . . . 64
- Interfaces lógicas . . . 64
- Configuração TCP/IP e ethernet . . . 66
- Configuração de FC SAN . . . 68
- Rede de conexão direta . . . 68

Configuração de storage em sistemas AFF/FAS . . . 69
- FC SAN . . . 69
- NFS . . . 74
- NVFAIL . . . 87
- Utilitário de reclinação ASMRU (ASMRU) . . . 88

Configuração de storage em sistemas ASA R2 . . . 88
- FC SAN . . . 88
- NVFAIL . . . 95
- Utilitário de recuperação ASM (ASRU) . . . 96

Virtualização . . . 97
- Capacidade de suporte . . . 97
- Apresentação de armazenamento . . . 97
- Drivers paravirtualizados . . . 98
- Sobrecarga da RAM . . . 99
- Striping do datastore . . . 99

Disposição em camadas . . . 100
- Visão geral . . . 100
- Políticas de disposição em camadas . . . 102
- Estratégias de disposição em camadas . . . 104
- Interrupções de acesso ao armazenamento de objetos . . . 108

Proteção de dados Oracle . . . 109
- Proteção de dados com o ONTAP . . . 109
- Planejamento de rto, RPO e SLA . . . 110
- Disponibilidade do banco de dados . . . 112
- Somas de verificação e integridade de dados . . . 114
- Noções básicas de backup e recuperação . . . 119

Recuperação de desastres Oracle . . . 133
- Visão geral . . . 133
- MetroCluster . . . 134
- Sincronização ativa do SnapMirror . . . 153

Migração de banco de dados Oracle . . . 188
- Visão geral . . . 188
- Planejamento de migração . . . 189
- Procedimentos . . . 192
- Exemplos de scripts . . . 298

Notas adicionais . . . 311

Otimização de desempenho e benchmarking 311
Stale NFSv3 fechaduras 314
Verificação do alinhamento do WAFL 315

Banco de dados Oracle

Bancos de dados Oracle no ONTAP

O ONTAP foi projetado para bancos de dados Oracle. Por décadas, o ONTAP foi otimizado para as demandas exclusivas de e/S de banco de dados relacional e vários recursos do ONTAP foram criados especificamente para atender às necessidades dos bancos de dados Oracle e até mesmo a pedido da própria Oracle Inc..

> Esta documentação substitui esses relatórios técnicos publicados anteriormente *TR-3633: Bancos de dados Oracle no ONTAP; TR-4591: Proteção de dados Oracle: Backup, recuperação, replicação; TR-4592: Oracle no MetroCluster; e TR-4534: Migração de bancos de dados Oracle para sistemas de armazenamento NetApp*

Além das muitas maneiras possíveis pelas quais o ONTAP agrega valor ao seu ambiente de banco de dados, há também uma grande variedade de requisitos do usuário, incluindo tamanho do banco de dados, requisitos de desempenho e necessidades de proteção de dados. As implantações conhecidas do storage NetApp incluem tudo, desde um ambiente virtualizado de aproximadamente 6.000 bancos de dados executados no VMware ESX até um data warehouse de instância única atualmente dimensionado a 996TB TB e em crescimento. Como resultado, existem algumas práticas recomendadas claras para configurar um banco de dados Oracle no storage NetApp.

Os requisitos para operar um banco de dados Oracle no storage NetApp são abordados de duas maneiras. Em primeiro lugar, quando existir uma boa prática clara, ela será chamada especificamente. Em um alto nível, muitas considerações de design que devem ser abordadas pelos arquitetos de soluções de armazenamento Oracle com base em seus requisitos de negócios específicos serão explicadas.

Configuração do ONTAP em sistemas AFF/ FAS

RAID

RAID refere-se ao uso de redundância para proteger os dados contra a perda de uma unidade.

Ocasionalmente, surgem perguntas sobre os níveis de RAID na configuração do armazenamento NetApp usado para bancos de dados Oracle e outros aplicativos empresariais. Muitas práticas recomendadas legadas da Oracle em relação à configuração do storage array contêm avisos sobre o uso do espelhamento RAID e/ou a prevenção de certos tipos de RAID. Embora apresentem pontos válidos, essas fontes não se aplicam ao RAID 4 e às tecnologias NetApp RAID DP e RAID-teC usadas no ONTAP.

RAID 4, RAID 5, RAID 6, RAID DP e RAID-teC usam paridade para garantir que a falha da unidade não resulte na perda de dados. Essas opções de RAID oferecem uma utilização do storage muito melhor em comparação com o espelhamento, mas a maioria das implementações de RAID tem uma desvantagem que afeta as operações de gravação. A conclusão de uma operação de gravação em outras implementações de RAID pode exigir várias leituras de unidade para regenerar os dados de paridade, um processo comumente chamado de penalidade de RAID.

O ONTAP, no entanto, não incorre nessa penalidade de RAID. Isso ocorre devido à integração do NetApp WAFL (Write Anywhere File Layout) com a camada RAID. As operações de gravação são Unidas na RAM e preparadas como um stripe RAID completo, incluindo geração de paridade. O ONTAP não precisa executar

uma leitura para concluir uma gravação, o que significa que o ONTAP e o WAFL evitam a penalidade de RAID. A performance de operações críticas à latência, como refazer o log, não é impedida, e as gravações aleatórias de arquivos de dados não incorrem em nenhuma penalidade de RAID resultante da necessidade de gerar novamente a paridade.

Com relação à confiabilidade estatística, até mesmo o RAID DP oferece melhor proteção do que o espelhamento RAID. O principal problema é a demanda feita nas unidades durante uma reconstrução RAID. Com um conjunto RAID espelhado, o risco de perda de dados de uma unidade falhar durante a reconstrução para seu parceiro no conjunto RAID é muito maior do que o risco de uma falha de unidade tripla em um conjunto RAID DP.

Gerenciamento de capacidade

Gerenciar um banco de dados ou outra aplicação empresarial com storage empresarial previsível, gerenciável e de alta performance exige espaço livre nas unidades para gerenciamento de dados e metadados. A quantidade de espaço livre necessária depende do tipo de unidade usada e dos processos de negócios.

O espaço livre é definido como qualquer espaço que não é usado para dados reais e inclui espaço não alocado no próprio agregado e espaço não utilizado dentro dos volumes constituintes. O thin Provisioning também deve ser considerado. Por exemplo, um volume pode conter um LUN de 1TB, dos quais apenas 50% são utilizados por dados reais. Em um ambiente thin Provisioning, isso parece estar consumindo 500GBMB de espaço corretamente. No entanto, em um ambiente totalmente provisionado, a capacidade total de 1TB TB parece estar em uso. O 500GB de espaço não alocado está oculto. Este espaço não é utilizado pelos dados reais e, portanto, deve ser incluído no cálculo do espaço livre total.

As recomendações da NetApp para sistemas de storage usados para aplicações empresariais são as seguintes:

Agregados SSD, incluindo sistemas AFF

A NetApp recomenda um mínimo de 10% de espaço livre. Isso inclui todo o espaço não utilizado, incluindo espaço livre dentro do agregado ou de um volume e qualquer espaço livre alocado devido ao uso de provisionamento completo, mas não é usado por dados reais. O espaço lógico não é importante, a questão é quanto espaço físico livre real está disponível para armazenamento de dados.

A recomendação de 10% de espaço livre é muito conservadora. Agregados SSD podem dar suporte a workloads em níveis ainda mais altos de utilização sem afetar a performance. No entanto, à medida que a utilização do agregado aumenta, o risco de ficar sem espaço também aumenta se a utilização não for monitorada com cuidado. Além disso, ao executar um sistema com 99% de capacidade pode não incorrer em uma penalidade de desempenho, mas provavelmente incorreria em esforço de gerenciamento tentando impedi-lo de preencher completamente enquanto hardware adicional é solicitado, e pode levar algum tempo para adquirir e instalar unidades adicionais.

Agregados de HDD, incluindo agregados Flash Pool

A NetApp recomenda um mínimo de 15% de espaço livre quando as unidades giratórias são usadas. Isso inclui todo o espaço não utilizado, incluindo espaço livre dentro do agregado ou de um volume e qualquer espaço livre alocado devido ao uso de provisionamento completo, mas não é usado por dados reais. O desempenho será impactado à medida que o espaço livre se aproxima de 10%.

Máquinas virtuais de armazenamento

O gerenciamento do storage de banco de dados Oracle é centralizado em uma Storage Virtual Machine (SVM).

Um SVM, conhecido como vserver na CLI do ONTAP, é uma unidade funcional básica de storage e é útil comparar um SVM com um convidado em um servidor VMware ESX.

Quando instalado pela primeira vez, o ESX não tem recursos pré-configurados, como hospedar um sistema operacional convidado ou suportar um aplicativo de usuário final. É um recipiente vazio até que uma máquina virtual (VM) seja definida. ONTAP é semelhante. Quando o ONTAP é instalado pela primeira vez, ele não tem funcionalidades de fornecimento de dados até que um SVM seja criado. É a personalidade do SVM que define os serviços de dados.

Assim como em outros aspectos da arquitetura de storage, as melhores opções para SVM e design de interface lógica (LIF) dependem muito dos requisitos de dimensionamento e das necessidades empresariais.

SVMs

Não há prática recomendada oficial para provisionar SVMs para ONTAP. A abordagem certa depende dos requisitos de gerenciamento e segurança.

A maioria dos clientes opera um SVM principal para a maioria de seus requisitos diários, mas cria um pequeno número de SVMs para necessidades especiais. Por exemplo, você pode querer criar:

- SVM para um banco de dados de negócios essencial gerenciado por uma equipe de especialistas
- Um SVM para um grupo de desenvolvimento a quem controle administrativo completo foi dado para que eles possam gerenciar seu próprio storage de forma independente
- SVM para dados empresariais confidenciais, como recursos humanos ou dados de relatórios financeiros, em que a equipe administrativa deve ser limitada

Em um ambiente de alocação a vários clientes, os dados de cada locatário podem receber um SVM dedicado. O limite para o número de SVMs e LIFs por cluster, par de HA e nó depende do protocolo que está sendo usado, do modelo de nó e da versão do ONTAP. Consulte o "NetApp Hardware Universe" para estes limites.

Gerenciamento de desempenho com QoS ONTAP

O gerenciamento de vários bancos de dados Oracle com segurança e eficiência requer uma estratégia de QoS eficaz. O motivo são os recursos de performance cada vez maiores de um sistema de storage moderno.

Especificamente, o aumento da adoção do storage all-flash possibilitou a consolidação de workloads. Os storage arrays que dependem de Mídia giratória costumavam dar suporte a apenas um número limitado de workloads de e/S com uso intenso de e/S devido aos recursos limitados de IOPS da tecnologia de unidade de rotação mais antiga. Um ou dois bancos de dados altamente ativos saturariam as unidades subjacentes muito antes que as controladoras de storage atingissem seus limites. Isso mudou. A capacidade de desempenho de um número relativamente pequeno de unidades SSD pode saturar até mesmo os controladores de armazenamento mais poderosos. Isso significa que todos os recursos das controladoras podem ser aproveitados sem o medo de colapso repentino da performance na medida em que picos de latência de Mídia giratórios.

Como exemplo de referência, um simples sistema HA AFF A800 de dois nós é capaz de atender a até um milhão de IOPS aleatórios antes que a latência suba acima de um milissegundo. Espera-se que muito poucas

cargas de trabalho individuais alcancem tais níveis. A utilização total desse array de sistema AFF A800 envolverá o hospedar vários workloads e isso com segurança, garantindo a previsibilidade requer controles de QoS.

Há dois tipos de qualidade de serviço (QoS) no ONTAP: IOPS e largura de banda. Controles de QoS podem ser aplicados a SVMs, volumes, LUNs e arquivos.

QoS de IOPS

Um controle de QoS de IOPS é obviamente baseado no total de IOPS de um determinado recurso, mas há vários aspetos da QoS de IOPS que podem não ser intuitivos. Alguns clientes inicialmente ficaram intrigados com o aparente aumento da latência quando um limite de IOPS é atingido. O aumento da latência é o resultado natural da limitação do IOPS. Logicamente, ele funciona de forma semelhante a um sistema de token. Por exemplo, se um determinado volume contendo datafiles tiver um limite de 10K IOPS, cada I/o que chegar deve receber primeiro um token para continuar o processamento. Desde que não mais de 10K tokens tenham sido consumidos em um determinado segundo, nenhum atraso está presente. Se as operações de e/S precisarem esperar para receber o token, essa espera aparecerá como latência adicional. Quanto mais difícil uma carga de trabalho ultrapassar o limite de QoS, mais tempo cada IO deve esperar na fila para que sua vez seja processada, o que parece ao usuário como maior latência.

> Tenha cuidado ao aplicar controles de QoS a dados de log de transação/refazer de banco de dados. Embora as demandas de desempenho do Registro de refazer sejam normalmente muito, muito mais baixas do que datafiles, a atividade do log de refazer é bursty. O IO acontece em breves pulsos, e um limite de QoS que parece apropriado para os níveis médios de e/S refazer pode ser muito baixo para os requisitos reais. O resultado pode ser limitações graves de desempenho, pois a QoS interage com cada sequência de log de refazer. Em geral, refazer e arquivar o log não deve ser limitado pela QoS.

QoS de largura de banda

Nem todos os tamanhos de e/S são iguais. Por exemplo, um banco de dados pode estar executando um grande número de leituras de blocos pequenos, o que resultaria no limite de IOPS ser atingido, mas os bancos de dados também podem estar executando uma operação de verificação de tabela completa que consistiria em um número muito pequeno de leituras de blocos grandes, consumindo uma quantidade muito grande de largura de banda, mas relativamente poucos IOPS.

Da mesma forma, um ambiente VMware pode gerar um número muito alto de IOPS aleatório durante a inicialização, mas executaria um iOS menor, mas maior, durante um backup externo.

Às vezes, o gerenciamento eficaz da performance exige limites de QoS de IOPS ou largura de banda, ou até mesmo ambos.

QoS mínima/garantida

Muitos clientes procuram uma solução que inclua QoS garantida, que é mais difícil de alcançar do que parece e é potencialmente um desperdício. Por exemplo, a colocação de bancos de dados 10 com uma garantia de 10K IOPS exige o dimensionamento de um sistema para um cenário em que todos os bancos de dados 10 estejam sendo executados simultaneamente a 10K IOPS, totalizando 100KK.

A melhor utilização para controles mínimos de QoS é proteger workloads críticos. Por exemplo, considere um controlador ONTAP com um máximo de IOPS possível de 500K K e uma combinação de workloads de produção e desenvolvimento. Você deve aplicar políticas máximas de QoS a workloads de desenvolvimento para impedir que qualquer banco de dados monopolize o controlador. Em seguida, você aplicaria políticas mínimas de QoS a workloads de produção para garantir que eles sempre tenham o IOPS necessário

disponível quando necessário.

QoS adaptável

QoS adaptável refere-se ao recurso ONTAP em que o limite de QoS é baseado na capacidade do objeto de storage. Raramente é usado com bancos de dados porque geralmente não há nenhum link entre o tamanho de um banco de dados e seus requisitos de desempenho. Bancos de dados grandes podem ser quase inertes, enquanto bancos de dados menores podem ser os mais intensivos em IOPS.

A QoS adaptável pode ser muito útil com armazenamentos de dados de virtualização, pois os requisitos de IOPS desses conjuntos de dados tendem a se correlacionar com o tamanho total do banco de dados. Um datastore mais recente contendo 1TB TB de arquivos VMDK provavelmente precisará de cerca de metade do desempenho como um datastore 2TB. A QoS adaptável permite que você aumente os limites de QoS automaticamente à medida que o armazenamento de dados se torna preenchido com dados.

Eficiência

Os recursos de eficiência de espaço do ONTAP são otimizados para bancos de dados Oracle. Em quase todos os casos, a melhor abordagem é deixar os padrões em vigor com todos os recursos de eficiência habilitados.

Os recursos de eficiência de espaço, como compressão, compactação e deduplicação, foram projetados para aumentar a quantidade de dados lógicos compatíveis com uma determinada quantidade de storage físico. O resultado são custos mais baixos e sobrecarga de gerenciamento.

Em um alto nível, a compressão é um processo matemático pelo qual padrões nos dados são detetados e codificados de uma forma que reduz os requisitos de espaço. Em contraste, a deduplicação deteta blocos repetidos reais de dados e remove as cópias externas. A compactação permite que vários blocos lógicos de dados compartilhem o mesmo bloco físico na Mídia.

> Veja as seções abaixo sobre provisionamento de thin para uma explicação da interação entre eficiência de armazenamento e reserva fracionária.

Compactação

Antes da disponibilidade de sistemas de storage all-flash, a compactação baseada em array era de valor limitado porque a maioria dos workloads com uso intenso de e/S exigia um número muito grande de fusos para fornecer desempenho aceitável. Os sistemas de armazenamento invariavelmente continham muito mais capacidade do que o necessário como um efeito colateral do grande número de unidades. A situação mudou com o aumento do armazenamento de estado sólido. Não há mais a necessidade de superprovisionamento vastamente de unidades puramente para obter um bom desempenho. O espaço da unidade em um sistema de armazenamento pode ser correspondente às necessidades reais de capacidade.

A maior funcionalidade de IOPS das solid-State drives (SSDs) quase sempre produz economias de custo em comparação com as unidades giratórias, mas a compactação pode gerar mais economias ao aumentar a capacidade efetiva da Mídia de estado sólido.

Existem várias maneiras de compactar dados. Muitos bancos de dados incluem suas próprias capacidades de compressão, mas isso raramente é observado em ambientes de clientes. A razão é geralmente a penalidade de desempenho para uma **mudança** para dados compatados, além disso, com alguns aplicativos, há altos custos de licenciamento para compactação no nível do banco de dados. Finalmente, há as consequências gerais de desempenho para as operações do banco de dados. Faz pouco sentido pagar um alto custo de licença por CPU para uma CPU que executa compactação de dados e descompressão em vez de trabalho

real de banco de dados. Uma opção melhor é descarregar o trabalho de compressão para o sistema de storage.

Compressão adaptável

A compactação adaptável foi totalmente testada com workloads empresariais sem efeito observado no desempenho, mesmo em um ambiente all-flash em que a latência é medida em microssegundos. Alguns clientes até relataram um aumento de desempenho com o uso de compactação porque os dados permanecem compatados no cache, aumentando efetivamente a quantidade de cache disponível em um controlador.

O ONTAP gerencia blocos físicos em 4KB unidades. A compactação adaptável usa um tamanho de bloco de compressão padrão de 8KB, o que significa que os dados são compatados em unidades 8KB. Isso corresponde ao tamanho do bloco 8KB mais frequentemente usado por bancos de dados relacionais. Os algoritmos de compressão tornam-se mais eficientes à medida que mais dados são compatados como uma única unidade. Um tamanho de bloco de compressão 32KB seria mais eficiente em termos de espaço do que uma unidade de bloco de compressão 8KB. Isso significa que a compactação adaptável usando o tamanho padrão de bloco 8KB leva a taxas de eficiência ligeiramente menores, mas também há um benefício significativo para o uso de um tamanho menor de bloco de compressão. As cargas de trabalho de banco de dados incluem uma grande quantidade de atividade de substituição. Substituir um 8KB de um bloco de dados comprimido 32KB requer a leitura de todo o 32KB de dados lógicos, descomprimindo-o, atualizando a região 8KB necessária, recomprimindo e gravando o 32KB inteiro de volta às unidades. Esta é uma operação muito cara para um sistema de storage e é o motivo pelo qual alguns storage arrays concorrentes baseados em tamanhos de bloco de compressão maiores também incorrer em uma penalidade de desempenho significativa nos workloads de banco de dados.

O tamanho do bloco usado pela compressão adaptável pode ser aumentado até 32KBMB. Isso pode melhorar a eficiência do storage e deve ser considerado para arquivos inativos, como logs de transações e arquivos de backup, quando uma quantidade substancial de tais dados é armazenada no array. Em algumas situações, os bancos de dados ativos que usam um tamanho de bloco 16KB ou 32KB também podem se beneficiar do aumento do tamanho de bloco da compactação adaptável a corresponder. Consulte um representante da NetApp ou do parceiro para obter orientação sobre se isso é apropriado para sua carga de trabalho.

Tamanhos de blocos de compactação maiores que 8KBMB não devem ser usados juntamente com a deduplicação em destinos de backup de streaming. A razão é que pequenas alterações nos dados de backup afetam a janela de compressão 32KB. Se a janela mudar, os dados compatados resultantes diferem em todo o arquivo. A deduplicação ocorre após a compactação, o que significa que o mecanismo de deduplicação vê cada backup compactado de forma diferente. Se a deduplicação de backups de streaming for necessária, somente a compactação adaptável de bloco 8KB deve ser usada. A compactação adaptável é preferível, porque funciona em um tamanho de bloco menor e não prejudica a eficiência da deduplicação. Por razões semelhantes, a compactação no lado do host também interfere na eficiência da deduplicação.

Alinhamento da compressão

A compactação adaptável em um ambiente de banco de dados requer alguma consideração do alinhamento do bloco de compressão. Fazer isso é apenas uma preocupação para os dados que estão sujeitos a substituições aleatórias de blocos muito específicos. Essa abordagem é semelhante em conceito ao alinhamento geral do sistema de arquivos, onde o início de um sistema de arquivos deve ser alinhado a um limite de dispositivo 4K e o tamanho de bloco de um sistema de arquivos deve ser um múltiplo de 4K.

Por exemplo, uma gravação 8KBD em um arquivo é compactada somente se ele se alinhar com um limite

8KBD dentro do próprio sistema de arquivos. Este ponto significa que ele deve cair no primeiro 8KB do arquivo, o segundo 8KB do arquivo, e assim por diante. A maneira mais simples de garantir o alinhamento correto é usar o tipo de LUN correto, qualquer partição criada deve ter um deslocamento desde o início do dispositivo que é um múltiplo de 8K, e usar um tamanho de bloco de sistema de arquivos que é um múltiplo do tamanho do bloco de banco de dados.

Dados como backups ou logs de transações são operações sequencialmente escritas que abrangem vários blocos, todos eles compatados. Portanto, não há necessidade de considerar o alinhamento. O único padrão de e/S de preocupação é as substituições aleatórias de arquivos.

Compactação de dados

A compactação de dados é uma tecnologia que melhora a eficiência da compressão. Como dito anteriormente, a compactação adaptável por si só pode proporcionar, na melhor das hipóteses, economias de 2:1x porque está limitada a armazenar uma e/S de 8KBx em um bloco de 4KB WAFL. Os métodos de compressão com tamanhos de bloco maiores proporcionam melhor eficiência. No entanto, eles não são adequados para dados sujeitos a pequenas substituições de blocos. A descompressão de 32KB unidades de dados, a atualização de uma porção 8KB, a recompressão e a gravação de volta nas unidades cria sobrecarga.

A compactação de dados funciona permitindo que vários blocos lógicos sejam armazenados em blocos físicos. Por exemplo, um banco de dados com dados altamente compressíveis, como texto ou blocos parcialmente completos, pode compactar de 8KB a 1KB. Sem compactação, 1KB PB de dados ainda ocupariam um bloco inteiro de 4KB TB. A compactação de dados in-line permite que 1KB TB de dados compatados sejam armazenados em apenas 1KB TB de espaço físico, juntamente com outros dados compatados. Não é uma tecnologia de compressão; é simplesmente uma maneira mais eficiente de alocar espaço nas unidades e, portanto, não deve criar qualquer efeito de desempenho detetável.

O grau de poupança obtido varia. Os dados que já estão compatados ou criptografados geralmente não podem ser mais compatados e, portanto, esses conjuntos de dados não se beneficiam com a compactação. Em contraste, datafiles recém-inicializados que contêm pouco mais do que metadados de bloco e zeros compactam até 80:1.

Eficiência de armazenamento sensível à temperatura

A eficiência de armazenamento sensível à temperatura (TSSE) está disponível no ONTAP 9.8 e posterior. Ele depende de mapas de calor de acesso a blocos para identificar blocos acessados com pouca frequência e compactá-los com maior eficiência.

Deduplicação

A deduplicação é a remoção de tamanhos de bloco duplicados de um conjunto de dados. Por exemplo, se o mesmo bloco 4KB existisse em 10 arquivos diferentes, a deduplicação redirecionaria esse bloco 4KB dentro de todos os arquivos 10 para o mesmo bloco físico 4KB. O resultado seria uma melhoria de 10:1 na eficiência para esses dados.

Dados como LUNs de inicialização convidado da VMware geralmente deduplicam muito bem porque consistem em várias cópias dos mesmos arquivos do sistema operacional. A eficiência de 100:1 e maior foi observada.

Alguns dados não contêm dados duplicados. Por exemplo, um bloco Oracle contém um cabeçalho que é globalmente exclusivo para o banco de dados e um trailer que é quase único. Como resultado, a deduplicação de um banco de dados Oracle raramente oferece mais de 1% de economia. A deduplicação com bancos de dados MS SQL é um pouco melhor, mas metadados exclusivos no nível de bloco ainda são uma limitação.

Economia de espaço de até 15% em bancos de dados com 16KB e grandes blocos foram observadas em alguns casos. O 4KB inicial de cada bloco contém o cabeçalho globalmente exclusivo, e o último bloco de 4KB contém o trailer quase único. Os blocos internos são candidatos à deduplicação, embora na prática isso seja quase inteiramente atribuído à deduplicação de dados zerados.

Muitos arrays concorrentes afirmam a capacidade de deduplicar bancos de dados com base na presunção de que um banco de dados é copiado várias vezes. A esse respeito, a deduplicação NetApp também pode ser usada, mas o ONTAP oferece uma opção melhor: A tecnologia NetApp FlexClone. O resultado final é o mesmo; várias cópias de um banco de dados que compartilham a maioria dos blocos físicos subjacentes são criadas. Usar o FlexClone é muito mais eficiente do que ter tempo para copiar arquivos de banco de dados e, em seguida, deduplicá-los. É, na verdade, não duplicação em vez de deduplicação, porque uma duplicata nunca é criada em primeiro lugar.

Eficiência e thin Provisioning

Os recursos de eficiência são formas de thin Provisioning. Por exemplo, um LUN de 100GB GB ocupando um volume de 100GB TB pode ser compactado para 50GB TB. Ainda não há economias reais realizadas porque o volume ainda é 100GB. O volume deve primeiro ser reduzido em tamanho para que o espaço guardado possa ser utilizado noutro local do sistema. Se as alterações posteriores ao LUN 100GBD resultarem em menos compressíveis os dados, o LUN crescerá em tamanho e o volume poderá ser preenchido.

O thin Provisioning é altamente recomendado porque pode simplificar o gerenciamento e fornecer melhorias substanciais na capacidade utilizável com economias de custo associadas. O motivo é simples: Os ambientes de banco de dados geralmente incluem muito espaço vazio, um grande número de volumes e LUNs e dados compressíveis. O provisionamento thick resulta na reserva de espaço no storage para volumes e LUNs, caso eles se tornem 100% cheios e contenham dados 100% não compactáveis. É pouco provável que isso ocorra. O thin Provisioning permite que esse espaço seja recuperado e usado em outros lugares e permite que o gerenciamento de capacidade seja baseado no próprio sistema de storage, em vez de muitos volumes e LUNs menores.

Alguns clientes preferem usar o provisionamento thick, seja para cargas de trabalho específicas ou, geralmente, com base em práticas operacionais e de aquisição estabelecidas.

> Se um volume for provisionado de forma grossa, deve-se ter cuidado para desativar completamente todos os recursos de eficiência para esse volume, incluindo descompressão e remoção de deduplicação usando `sis undo` o comando. O volume não deve aparecer `volume efficiency show` na saída. Se isso acontecer, o volume ainda será parcialmente configurado para recursos de eficiência. Como resultado, as garantias de substituição funcionam de forma diferente, o que aumenta a chance de que a configuração seja ultrapassada fazendo com que o volume fique inesperadamente sem espaço, resultando em erros de e/S do banco de dados.

Práticas recomendadas de eficiência

A NetApp recomenda o seguinte:

Padrões do AFF

Os volumes criados no ONTAP executados em um sistema all-flash AFF são thin Provisioning com todos os recursos de eficiência in-line habilitados. Embora os bancos de dados geralmente não se beneficiem da deduplicação e possam incluir dados não compressíveis, as configurações padrão são, no entanto, apropriadas para quase todas as cargas de trabalho. O ONTAP foi projetado para processar com eficiência todos os tipos de dados e padrões de e/S, resultando ou não em economia. Os padrões só devem ser alterados se os motivos forem totalmente compreendidos e houver um benefício para se desviar.

Recomendações gerais

- Se os volumes e/ou LUNs não estiverem provisionados de forma fina, você deverá desativar todas as configurações de eficiência porque o uso desses recursos não oferece economia e a combinação de provisionamento espesso com eficiência de espaço habilitada pode causar comportamento inesperado, incluindo erros fora do espaço.
- Se os dados não estiverem sujeitos a sobrescritas, como backups ou logs de transação de banco de dados, você poderá obter maior eficiência ativando o TSSE com um período de resfriamento baixo.
- Alguns arquivos podem conter uma quantidade significativa de dados não compressíveis, por exemplo, quando a compactação já está ativada no nível de aplicativo de arquivos são criptografados. Se qualquer um desses cenários for verdadeiro, considere desativar a compactação para permitir uma operação mais eficiente em outros volumes que contenham dados compressíveis.
- Não use a compactação e a deduplicação do 32KB com backups de bancos de dados. Consulte a secção Compressão adaptável para obter detalhes.

Thin Provisioning

O thin Provisioning para um banco de dados Oracle requer um Planejamento cuidadoso, pois o resultado é configurar mais espaço em um sistema de armazenamento do que necessariamente está fisicamente disponível. Vale muito a pena o esforço porque, quando feito corretamente, o resultado é uma significativa economia de custos e melhorias na capacidade de gerenciamento.

O thin Provisioning vem em várias formas e é parte integrante de muitos recursos que a ONTAP oferece a um ambiente de aplicações empresariais. O provisionamento de thin também está intimamente relacionado às tecnologias de eficiência pelo mesmo motivo: Os recursos de eficiência permitem que mais dados lógicos sejam armazenados do que o que existe tecnicamente no sistema de storage.

Quase qualquer uso de snapshots envolve thin Provisioning. Por exemplo, um banco de dados 10TB típico no storage NetApp inclui cerca de 30 dias de snapshots. Esse arranjo resulta em aproximadamente 10TB GB de dados visíveis no sistema de arquivos ativo e 300TB dedicados a snapshots. O total de 310TB TB de storage geralmente reside em aproximadamente 12TB a 15TB TB de espaço. O banco de dados ativo consome 10TB, e os 300TB restantes de dados requerem apenas 2TB a 5TB GB de espaço porque apenas as alterações aos dados originais são armazenadas.

A clonagem também é um exemplo de thin Provisioning. Um grande cliente da NetApp criou 40 clones de um banco de dados 80TB para uso por desenvolvimento. Se todos os desenvolvedores do 40 que usam esses clones sobrescrevessem cada bloco em cada arquivo de dados, mais de 3,2PB GB de armazenamento seriam necessários. Na prática, a rotatividade é baixa e a exigência de espaço coletivo está mais próxima de 40TB, porque apenas as alterações são armazenadas nas unidades.

Gerenciamento de espaço

Alguns cuidados devem ser tomados com o thin Provisioning de um ambiente de aplicativo porque as taxas de alteração de dados podem aumentar inesperadamente. Por exemplo, o consumo de espaço devido a snapshots pode crescer rapidamente se as tabelas do banco de dados forem reindexadas ou se aplicar patches em larga escala aos convidados VMware. Um backup perdido pode gravar uma grande quantidade de dados em um tempo muito curto. Finalmente, pode ser difícil recuperar alguns aplicativos se um sistema de arquivos ficar sem espaço livre inesperadamente.

Felizmente, esses riscos podem ser resolvidos com uma configuração cuidadosa de `volume-autogrow` políticas e `snapshot-autodelete`. Como os nomes deles implicam, essas opções permitem que um

usuário crie políticas que desobstruam automaticamente o espaço consumido por snapshots ou aumentem um volume para acomodar dados adicionais. Muitas opções estão disponíveis e as necessidades variam de acordo com o cliente.

Consulte o "documentação de gerenciamento de storage lógico" para obter uma discussão completa sobre esses recursos.

Reservas fracionárias

A reserva fracionária refere-se ao comportamento de um LUN em um volume com relação à eficiência do espaço. Quando a opção `fractional-reserve` é definida para 100%, todos os dados no volume podem experimentar 100% de rotatividade com qualquer padrão de dados sem esgotar espaço no volume.

Por exemplo, considere um banco de dados em um único LUN 250GB em um volume 1TB. Criar um instantâneo resultaria imediatamente na reserva de 250GBMB de espaço adicional no volume para garantir que o volume não fique sem espaço por qualquer motivo. O uso de reservas fracionárias geralmente é desperdiçado porque é extremamente improvável que cada byte no volume do banco de dados precise ser substituído. Não há razão para reservar espaço para um evento que nunca acontece. Ainda assim, se um cliente não puder monitorar o consumo de espaço em um sistema de armazenamento e tiver certeza de que o espaço nunca se esgote, reservas fracionárias de 100% seriam necessárias para usar snapshots.

Compactação e deduplicação

Compactação e deduplicação são ambas formas de thin Provisioning. Por exemplo, um espaço físico de dados de 50TB GB pode ser compactado para 30TB GB, resultando em uma economia de 20TB GB. Para que a compactação possa gerar quaisquer benefícios, alguns desses 20TB precisam ser usados para outros dados, ou o sistema de storage precisa ser adquirido com menos de 50TB TB. O resultado é o armazenamento de mais dados do que o tecnicamente disponível no sistema de armazenamento. Do ponto de vista dos dados, há 50TB TB de dados, embora ocupe apenas 30TB TB nas unidades.

Há sempre a possibilidade de que a compressibilidade de um conjunto de dados mude, o que resultaria em aumento do consumo de espaço real. Esse aumento no consumo significa que a compactação deve ser gerenciada como com outras formas de thin Provisioning em termos de monitoramento e uso `volume-autogrow` e `snapshot-autodelete`.

A compactação e a deduplicação são discutidas em mais detalhes na seção xref:./oracle/efficiency.html

Compressão e reservas fracionárias

A compactação é uma forma de thin Provisioning. Reservas fracionárias afetam o uso da compressão, com uma nota importante; o espaço é reservado antes da criação do snapshot. Normalmente, a reserva fracionária só é importante se existir um instantâneo. Se não houver instantâneo, a reserva fracionária não é importante. Este não é o caso da compressão. Se um LUN for criado em um volume com compactação, o ONTAP preservará espaço para acomodar um snapshot. Esse comportamento pode ser confuso durante a configuração, mas é esperado.

Por exemplo, considere um volume de 10GB TB com um LUN de 5GB GB que foi comprimido até 2,5GB TB sem instantâneos. Considere estes dois cenários:

- A reserva fracionária é de 100 resultados na utilização de 7,5GB
- A reserva fracionária é de 0 resultados na utilização de 2,5GB

O primeiro cenário inclui 2,5GB GB de consumo de espaço para dados atuais e 5GB GB de espaço para representar 100% de rotatividade da fonte em antecipação ao uso de instantâneos. O segundo cenário não

reserva espaço extra.

Embora esta situação possa parecer confusa, é improvável que seja encontrada na prática. A compactação implica thin Provisioning, e thin Provisioning em um ambiente LUN requer reservas fracionárias. É sempre possível que os dados compatados sejam sobrescritos por algo não compressível, o que significa que um volume deve ser thin Provisioning para compactação para resultar em qualquer economia.

A NetApp recomenda as seguintes configurações de reserva:

- Defina `fractional-reserve` como 0 quando o monitoramento de capacidade básica estiver em vigor, juntamente com `volume-autogrow` e `snapshot-autodelete`.
- Defina `fractional-reserve` para 100 se não houver capacidade de monitorização ou se for impossível esgotar o espaço em qualquer circunstância.

Espaço livre e alocação de espaço LVM

A eficiência do provisionamento dinâmico de LUNs ativos em um ambiente de sistema de arquivos pode ser perdida ao longo do tempo à medida que os dados são excluídos. A menos que os dados excluídos sejam sobrescritos com zeros (veja também ASMRU ou o espaço seja liberado com a recuperação de espaço TRIM/UNMAP, os dados "apagados" ocupam cada vez mais espaço em branco não alocado no sistema de arquivos. Além disso, o provisionamento dinâmico de LUNs ativos tem utilidade limitada em muitos ambientes de banco de dados, porque os arquivos de dados são inicializados com seu tamanho total no momento da criação.

Um Planejamento cuidadoso da configuração do LVM pode melhorar a eficiência e minimizar a necessidade de provisionamento de armazenamento e redimensionamento de LUN. Quando um LVM, como Veritas VxVM ou Oracle ASM, é usado, os LUNs subjacentes são divididos em extensões que só são usadas quando necessário. Por exemplo, se um conjunto de dados começar com 2TB TB de tamanho, mas puder crescer para 10TB TB com o tempo, esse conjunto de dados poderá ser colocado em 10TB PB de LUNs provisionados, organizados em um grupo de discos LVM. Ele ocuparia apenas 2TBMB de espaço no momento da criação e só reivindicaria espaço adicional, pois as extensões são alocadas para acomodar o crescimento dos dados. Este processo é seguro, desde que o espaço seja monitorado.

Failover/switchover da ONTAP

É necessário entender as funções de takeover e switchover do storage para garantir que as operações do banco de dados Oracle não sejam interrompidas por essas operações. Além disso, os argumentos usados pelas operações de aquisição e comutação podem afetar a integridade dos dados se usados incorretamente.

- Em condições normais, as gravações recebidas em um determinado controlador são espelhadas de forma síncrona para seu parceiro. Em um ambiente NetApp MetroCluster, as gravações também são espelhadas em um controle remoto. Até que uma gravação seja armazenada em Mídia não volátil em todos os locais, ela não será reconhecida para a aplicação host.
- A Mídia que armazena os dados de gravação é chamada de memória não volátil ou NVMEM. Ele também é por vezes referido à memória de acesso aleatório não volátil (NVRAM), e pode ser considerado como um cache de escrita, embora funcione como um diário. Em uma operação normal, os dados do NVMEM não são lidos; eles são usados apenas para proteger os dados em caso de falha de software ou hardware. Quando os dados são gravados em unidades, os dados são transferidos da RAM no sistema, e não da NVMEM.
- Durante uma operação de takeover, um nó em um par de alta disponibilidade (HA) assume as operações

de seu parceiro. Um switchover é essencialmente o mesmo, mas se aplica às configurações do MetroCluster nas quais um nó remoto assume as funções de um nó local.

Durante as operações de manutenção de rotina, uma operação de aquisição ou comutação de storage deve ser transparente, exceto para uma breve pausa em operações à medida que os caminhos de rede mudam. No entanto, a rede pode ser complicada, e é fácil fazer erros. Por isso, a NetApp recomenda fortemente testar cuidadosamente as operações de takeover e switchover antes de colocar um sistema de storage em produção. Fazer isso é a única maneira de ter certeza de que todos os caminhos de rede estão configurados corretamente. Em um ambiente SAN, verifique cuidadosamente a saída do comando `sanlun lun show -p` para garantir que todos os caminhos primários e secundários esperados estejam disponíveis.

Deve ser tomado cuidado ao emitir uma tomada ou mudança forçada. Forçar uma alteração na configuração de armazenamento com essas opções significa que o estado do controlador que possui as unidades é ignorado e o nó alternativo assume forçosamente o controle das unidades. Forçar uma aquisição incorreta pode resultar em perda ou corrupção de dados. Isso ocorre porque uma tomada ou comutação forçada pode descartar o conteúdo do NVMEM. Após a conclusão do takeover ou switchover, a perda desses dados significa que os dados armazenados nas unidades podem reverter para um estado um pouco mais antigo do ponto de vista do banco de dados.

Uma aquisição forçada com um par de HA normal raramente deve ser necessária. Em quase todos os cenários de falha, um nó desliga e informa o parceiro para que ocorra um failover automático. Existem alguns casos de borda, como uma falha contínua em que a interconexão entre nós é perdida e, em seguida, um controlador é perdido, na qual uma tomada pública forçada é necessária. Em tal situação, o espelhamento entre nós é perdido antes da falha do controlador, o que significa que o controlador sobrevivente não teria mais uma cópia das gravações em andamento. Então, a aquisição precisa ser forçada, o que significa que os dados podem ser perdidos.

A mesma lógica se aplica a um switchover MetroCluster. Em condições normais, uma mudança é quase transparente. No entanto, um desastre pode resultar em uma perda de conetividade entre o local sobrevivente e o local do desastre. Do ponto de vista do site sobrevivente, o problema não poderia ser mais do que uma interrupção na conetividade entre sites, e o site original ainda pode estar processando dados. Se um nó não puder verificar o estado do controlador principal, somente um switchover forçado é possível.

A NetApp recomenda tomar as seguintes precauções:

- Tenha muito cuidado para não forçar acidentalmente uma aquisição ou uma mudança. Normalmente, forçar não deve ser necessário e forçar a mudança pode causar perda de dados.
- Se for necessário um controle ou switchover forçado, certifique-se de que os aplicativos sejam desligados, todos os sistemas de arquivos sejam desmontados e os grupos de volume do gerenciador de volumes lógico (LVM) sejam variados. Os grupos de discos ASM devem ser desmontados.
- No caso de um switchover forçado do MetroCluster, feche o nó com falha de todos os recursos de storage sobreviventes. Para obter mais informações, consulte o Guia de gerenciamento e recuperação de desastres do MetroCluster para obter a versão relevante do ONTAP.

MetroCluster e vários agregados

O MetroCluster é uma tecnologia de replicação síncrona que alterna para o modo assíncrono se a conetividade for interrompida. Essa é a solicitação mais comum dos clientes, pois a replicação síncrona garantida significa que a interrupção na conectividade do local leva a uma interrupção completa da e/S do banco de dados, retirando o banco de dados do serviço.

Com o MetroCluster, agregados ressincronizam rapidamente após a restauração da conectividade. Diferentemente de outras tecnologias de storage, o MetroCluster nunca deve exigir um reespelhamento completo após falha do local. Apenas as alterações delta devem ser enviadas.

Em conjuntos de dados que abrangem agregados, existe um pequeno risco de que etapas adicionais de recuperação de dados sejam necessárias em um cenário de desastre contínuo. Especificamente, se (a) a conetividade entre sites for interrompida, (b) a conetividade for restaurada, (c) os agregados atingirem um estado em que alguns são sincronizados e alguns não são, e então (d) o local primário é perdido, o resultado é um local sobrevivente no qual os agregados não são sincronizados uns com os outros. Se isso acontecer, partes do conjunto de dados são sincronizadas entre si e não é possível abrir aplicativos, bancos de dados ou datastores sem recuperação. Se um conjunto de dados abranger agregados, o NetApp recomenda fortemente a utilização de backups baseados em snapshot com uma das muitas ferramentas disponíveis para verificar a capacidade de recuperação rápida neste cenário incomum.

Configuração do ONTAP em sistemas ASA r2

RAID

RAID refere-se ao uso de redundância baseada em paridade para proteger os dados contra falhas de disco. O ASA r2 utiliza as mesmas tecnologias ONTAP RAID dos sistemas AFF e FAS , garantindo uma proteção robusta contra falhas em múltiplos discos.

O ONTAP realiza a configuração RAID automaticamente para sistemas ASA r2. Este é um componente essencial da experiência simplificada de gerenciamento de armazenamento introduzida com a personalidade ASA r2.

Detalhes importantes sobre a configuração automática de RAID no ASA r2 incluem:

- Zonas de Disponibilidade de Armazenamento (SAZ): Em vez de gerenciar manualmente agregados e grupos RAID tradicionais, o ASA r2 usa Zonas de Disponibilidade de Armazenamento (SAZs). Tratam-se de conjuntos de discos compartilhados e protegidos por RAID para um par de alta disponibilidade (HA), onde ambos os nós têm acesso total ao mesmo armazenamento.
- Posicionamento automático: Quando uma unidade de armazenamento (LUN ou namespace NVMe) é criada, o ONTAP cria automaticamente um volume dentro da SAZ e o posiciona para otimizar o desempenho e o equilíbrio de capacidade.
- Sem gerenciamento manual de agregados: os comandos tradicionais de gerenciamento de agregados e grupos RAID não são suportados no ASA r2. Isso elimina a necessidade de os administradores planejarem manualmente os tamanhos dos grupos RAID, os discos de paridade ou as atribuições de nós.
- Provisionamento simplificado: O provisionamento é gerenciado por meio do System Manager ou de comandos CLI simplificados que se concentram nas unidades de armazenamento em vez do layout RAID físico subjacente.
- Rebalanceamento de carga de trabalho: A partir das versões de 2025 (ONTAP 9.17.1), o ONTAP rebalanceia automaticamente as cargas de trabalho entre os nós do par HA para garantir que o desempenho e a utilização do espaço permaneçam equilibrados sem intervenção manual.

O ASA r2 utiliza automaticamente as tecnologias RAID padrão do ONTAP: RAID DP para a maioria das configurações e RAID-TEC para pools de SSDs muito grandes. Isso elimina a necessidade de seleção manual de RAID. Esses níveis de RAID baseados em paridade oferecem melhor eficiência e confiabilidade de armazenamento do que o espelhamento, que as práticas recomendadas mais antigas da Oracle costumam sugerir, mas não são relevantes para o ASA r2. O ONTAP evita a penalidade de gravação RAID tradicional por

meio da integração com o WAFL , garantindo desempenho ideal para cargas de trabalho do Oracle, como redo logging e gravações aleatórias de arquivos de dados. Combinado com o gerenciamento automatizado de RAID e as Zonas de Disponibilidade de Armazenamento (SAZs), o ASA r2 oferece alta disponibilidade e proteção de nível empresarial para bancos de dados Oracle.

Gerenciamento de capacidade

Gerenciar um banco de dados ou outra aplicação empresarial com storage empresarial previsível, gerenciável e de alta performance exige espaço livre nas unidades para gerenciamento de dados e metadados. A quantidade de espaço livre necessária depende do tipo de unidade usada e dos processos de negócios.

O ASA r2 usa Zonas de Disponibilidade de Armazenamento (SAZ) em vez de agregados, mas o princípio permanece o mesmo: o espaço livre inclui qualquer capacidade física não consumida por dados reais, snapshots ou sobrecarga do sistema. O provisionamento dinâmico também deve ser considerado — as alocações lógicas não refletem o uso físico real.

As recomendações da NetApp para sistemas de armazenamento ASA r2 usados em aplicações empresariais são as seguintes:

Pools de SSD em sistemas ASA r2

* A NetApp recomenda* manter um mínimo de 10% de espaço físico livre em ambientes ASA r2. Esta diretriz se aplica a pools compostos exclusivamente por SSDs usados por sistemas ASA r2 e inclui todo o espaço não utilizado dentro da SAZ e das unidades de armazenamento. O espaço lógico é irrelevante; o foco está no espaço físico livre disponível para armazenamento de dados.

Embora o ASA r2 possa suportar alta utilização sem degradação de desempenho, operar próximo à capacidade máxima aumenta o risco de esgotamento do espaço e a sobrecarga administrativa ao expandir o armazenamento. Operar com mais de 90% de utilização pode não afetar o desempenho, mas pode complicar o gerenciamento e atrasar o provisionamento de unidades adicionais.

Os sistemas ASA r2 suportam unidades de armazenamento de até 128 TB e tamanhos de SAZ de até 2 PB por par HA, com o ONTAP balanceando automaticamente a capacidade entre os nós. O monitoramento da utilização nos níveis de cluster, SAZ e unidade de armazenamento é essencial para garantir espaço livre adequado para snapshots, cargas de trabalho com provisionamento dinâmico e crescimento futuro. Se a capacidade se aproximar de limites críticos (utilização em torno de 90%), SSDs adicionais devem ser adicionados em grupos (mínimo de seis unidades) para manter o desempenho e a resiliência.

Máquinas virtuais de armazenamento

O gerenciamento de armazenamento do banco de dados Oracle em sistemas ASA r2 também é centralizado em uma Máquina Virtual de Armazenamento (SVM), conhecida como vserver na CLI do ONTAP .

Uma SVM (Storage Virtual Machine) é a unidade fundamental de provisionamento e segurança de armazenamento no ONTAP, semelhante a uma VM convidada em um servidor VMware ESX. Quando o ONTAP é instalado pela primeira vez no ASA r2, ele não possui recursos de fornecimento de dados até que uma SVM seja criada. A SVM define a personalidade e os serviços de dados para o ambiente SAN.

Os sistemas ASA r2 utilizam uma personalidade ONTAP exclusiva para SAN, otimizada para suportar

protocolos de bloco (FC, iSCSI, NVMe/FC, NVMe/TCP) e que remove recursos relacionados a NAS. Isso simplifica o gerenciamento e garante que todas as configurações de SVM sejam otimizadas para cargas de trabalho SAN. Diferentemente dos sistemas AFF/ FAS , o ASA r2 não expõe opções para serviços NAS, como diretórios pessoais ou compartilhamentos NFS.

Quando um cluster é criado, o ASA r2 provisiona automaticamente uma SVM de dados padrão chamada svm1 com os protocolos SAN habilitados. Esta SVM está pronta para operações de armazenamento em bloco sem a necessidade de configuração manual dos serviços de protocolo. Por padrão, as LIFs de dados IP nesta SVM suportam os protocolos iSCSI e NVMe/TCP e utilizam a política de serviço default-data-blocks, o que simplifica a configuração inicial para cargas de trabalho SAN. Posteriormente, os administradores podem criar SVMs adicionais ou personalizar as configurações do LIF com base em requisitos de desempenho, segurança ou multilocação.

> As interfaces lógicas (LIFs) para protocolos SAN devem ser projetadas com base nos requisitos de desempenho e disponibilidade. O ASA r2 suporta LIFs iSCSI, FC e NVMe, mas observe que o failover automático de LIF iSCSI não está habilitado por padrão, pois o ASA r2 usa rede compartilhada para hosts NVMe e SCSI. Para habilitar o failover automático, crie "LIFs somente para iSCSI".

SVMs

Assim como em outras plataformas ONTAP , não existe uma prática recomendada oficial para o número de SVMs a serem criadas; a decisão depende dos requisitos de gerenciamento e segurança.

A maioria dos clientes opera uma única SVM principal para as operações diárias e cria SVMs adicionais para necessidades específicas, como:

- Uma SVM dedicada para um banco de dados crítico de negócios, gerenciada por uma equipe especializada.
- Uma SVM para um grupo de desenvolvimento com controle administrativo delegado.
- Uma SVM para dados sensíveis que exigem acesso administrativo restrito.

Em ambientes com múltiplos inquilinos, cada inquilino pode ter uma SVM dedicada atribuída. O limite para o número de SVMs e LIFs por cluster, par HA e nó depende do protocolo utilizado, do modelo do nó e da versão do ONTAP. Consulte o "NetApp Hardware Universe" para esses limites.

> O ASA r2 suporta até 256 SVMs por cluster e por par HA a partir do ONTAP 9.18.1 (anteriormente 32 em versões anteriores).

Gerenciamento de desempenho com ONTAP QoS em sistemas ASA r2

Gerenciar com segurança e eficiência vários bancos de dados Oracle no ASA r2 requer uma estratégia de QoS eficaz. Isso é especialmente importante porque os sistemas ASA r2 são plataformas SAN totalmente em flash, projetadas para desempenho extremamente alto e consolidação de cargas de trabalho.

Um número relativamente pequeno de SSDs pode saturar até mesmo os controladores mais potentes, portanto, os controles de QoS são essenciais para garantir um desempenho previsível em diversas cargas de trabalho. Como referência, os sistemas ASA r2, como o ASA A1K ou A90, podem fornecer centenas de milhares a mais de um milhão de IOPS com latência inferior a um milissegundo. Pouquíssimas cargas de trabalho individuais consumiriam esse nível de desempenho, portanto, a utilização plena normalmente envolve a hospedagem de vários bancos de dados ou aplicativos. Fazer isso com segurança requer políticas de QoS

para evitar a disputa por recursos.

O ONTAP QoS no ASA r2 funciona da mesma forma que nos sistemas AFF/ FAS , com dois tipos principais de controles: IOPS e largura de banda. Os controles de QoS podem ser aplicados a SVMs e LUNs.

QoS de IOPS

A QoS baseada em IOPS limita o total de IOPS para um determinado recurso. No ASA r2, as políticas de QoS podem ser aplicadas no nível do SVM e a objetos de armazenamento individuais, como LUNs. Quando uma carga de trabalho atinge seu limite de IOPS, solicitações adicionais de E/S são enfileiradas para obtenção de tokens, o que introduz latência. Esse é o comportamento esperado e impede que uma única carga de trabalho monopolize os recursos do sistema.

> Tenha cautela ao aplicar controles de QoS aos dados de log de transações/refazer do banco de dados. Essas cargas de trabalho são intermitentes, e um limite de QoS que parece razoável para a atividade média pode ser muito baixo para picos de atividade, causando sérios problemas de desempenho. Em geral, o registro de refazer e arquivar não deve ser limitado pela QoS.

QoS de largura de banda

A QoS baseada em largura de banda limita a taxa de transferência em Mbps. Isso é útil quando as cargas de trabalho executam leituras ou gravações de blocos grandes, como varreduras completas de tabelas ou operações de backup, que consomem uma largura de banda significativa, mas relativamente poucas operações de E/S (IOPS). A combinação de limites de IOPS e de largura de banda pode proporcionar um controle mais preciso.

QoS mínima/garantida

As políticas mínimas de QoS reservam o desempenho para cargas de trabalho críticas. Por exemplo, em um ambiente misto com bancos de dados de produção e desenvolvimento, aplique o máximo de QoS às cargas de trabalho de desenvolvimento e o mínimo de QoS às cargas de trabalho de produção para garantir um desempenho previsível.

QoS adaptável

O QoS adaptativo ajusta os limites com base no tamanho do objeto de armazenamento. Embora raramente utilizado para bancos de dados (porque o tamanho não se correlaciona com as necessidades de desempenho), pode ser útil para cargas de trabalho de virtualização, onde os requisitos de desempenho aumentam com a capacidade.

Eficiência

Os recursos de eficiência espacial do ONTAP são totalmente compatíveis e otimizados para sistemas ASA r2. Na maioria dos casos, a melhor abordagem é manter as configurações padrão com todos os recursos de eficiência ativados.

Os sistemas ASA r2 são plataformas SAN totalmente em flash, portanto, tecnologias de eficiência como compressão, compactação e deduplicação são essenciais para maximizar a capacidade utilizável e reduzir custos.

Compactação

A compressão reduz os requisitos de espaço ao codificar padrões nos dados. Com sistemas ASA r2 baseados em SSD, a compressão proporciona economias significativas, pois a memória flash elimina a necessidade de provisionamento excessivo para obter desempenho. A compressão adaptativa do ONTAP está ativada por padrão e foi exaustivamente testada com cargas de trabalho corporativas, incluindo bancos de dados Oracle, sem impacto mensurável no desempenho — mesmo em ambientes onde a latência é medida em microssegundos. Em alguns casos, o desempenho melhora porque os dados comprimidos ocupam menos espaço na cache.

> A eficiência de armazenamento sensível à temperatura (TSSE) não se aplica aos sistemas ASA r2. Nos sistemas ASA r2, a compressão não se baseia em dados "quentes" (acessados com frequência) ou dados "frios" (acessados com pouca frequência). A compressão começa sem esperar que os dados esfriem.

Compressão adaptável

A compressão adaptativa usa um tamanho de bloco de 8 KB por padrão, que corresponde ao tamanho de bloco comumente usado por bancos de dados relacionais. Tamanhos de bloco maiores (16 KB ou 32 KB) podem melhorar a eficiência para dados sequenciais, como logs de transações ou backups, mas devem ser usados com cautela em bancos de dados ativos para evitar sobrecarga durante as sobrescritas.

> O tamanho do bloco pode ser aumentado até 32 KB para arquivos inativos, como registros ou backups. Consulte as orientações da NetApp antes de alterar as configurações padrão.

> Não utilize compressão de 32 KB com deduplicação para backups de streaming. Use compressão de 8 KB para manter a eficiência da desduplicação.

Alinhamento da compressão

O alinhamento da compressão é importante para sobrescritas aleatórias. Certifique-se de que o tipo de LUN, o deslocamento da partição (múltiplo de 8 KB) e o tamanho do bloco do sistema de arquivos estejam alinhados ao tamanho do bloco do banco de dados. Dados sequenciais, como backups ou registros, não exigem considerações de alinhamento.

Compactação de dados

A compactação complementa a compressão, permitindo que vários blocos comprimidos compartilhem o mesmo bloco físico. Por exemplo, se um bloco de 8 KB for comprimido para 1 KB, a compactação garante que o espaço restante não seja desperdiçado. Essa funcionalidade é integrada ao código e não acarreta penalidades de desempenho.

Deduplicação

A desduplicação remove blocos duplicados em conjuntos de dados. Embora os bancos de dados Oracle normalmente apresentem economias mínimas de deduplicação devido aos cabeçalhos e rodapés de bloco exclusivos, a deduplicação do ONTAP ainda pode recuperar espaço de blocos zerados e padrões repetidos.

Eficiência e thin Provisioning

Os sistemas ASA r2 utilizam provisionamento dinâmico por padrão. Os recursos de eficiência complementam o provisionamento dinâmico para maximizar a capacidade utilizável.

Nos sistemas de armazenamento ASA r2, as unidades de armazenamento são sempre provisionadas de forma dinâmica. O provisionamento espesso não é suportado.

Tecnologia QuickAssist (QAT)

Nas plataformas NetApp ASA r2, a tecnologia Intel QuickAssist (QAT) oferece eficiência acelerada por hardware que difere significativamente da eficiência de armazenamento sensível à temperatura (TSSE) baseada em software, sem a QAT.

QAT com aceleração de hardware:

- Descarrega as tarefas de compressão e criptografia dos núcleos da CPU.
- Permite eficiência imediata e em linha tanto para dados quentes (acessados com frequência) quanto para dados frios (acessados com pouca frequência).
- Reduz significativamente a sobrecarga da CPU.
- Proporciona maior taxa de transferência e menor latência.
- Melhora a escalabilidade para operações que exigem alto desempenho, como criptografia TLS e VPN.

TSSE sem QAT:

- Depende de processos controlados pela CPU para operações eficientes.
- Aplica a eficiência somente a dados frios após um atraso.
- Consome mais recursos da CPU.
- Limita o desempenho geral em comparação com sistemas acelerados por QAT.

Os modernos sistemas ASA r2 oferecem, portanto, maior rapidez, eficiência acelerada por hardware e melhor utilização do sistema do que as plataformas mais antigas que utilizavam apenas TSSE.

Melhores práticas de eficiência para ASA r2

A NetApp recomenda o seguinte:

Configurações padrão do ASA r2

As unidades de armazenamento criadas no ONTAP em execução em sistemas ASA r2 são provisionadas dinamicamente com todos os recursos de eficiência inline ativados por padrão, incluindo compressão, compactação e deduplicação. Embora os bancos de dados Oracle geralmente não se beneficiem significativamente da deduplicação e possam incluir dados não compressíveis, essas configurações padrão são adequadas para quase todas as cargas de trabalho. O ONTAP foi projetado para processar com eficiência todos os tipos de dados e padrões de E/S, independentemente de resultarem ou não em economia de custos. Os valores padrão só devem ser alterados se os motivos forem totalmente compreendidos e houver um benefício claro em se desviar deles.

Recomendações gerais

- Desative a compressão para dados criptografados ou compactados pelo aplicativo: se os arquivos já estiverem compactados no nível do aplicativo ou criptografados, desative a compressão para otimizar o desempenho e permitir uma operação mais eficiente em outras unidades de armazenamento.
- Evite combinar blocos de compressão grandes com deduplicação: Não utilize compressão de 32 KB e deduplicação simultaneamente para backups de banco de dados. Para backups em fluxo contínuo, use compressão de 8 KB para manter a eficiência da deduplicação.

- Monitore a economia de espaço: utilize as ferramentas do ONTAP (System Manager, Active IQ) para acompanhar a economia de espaço real e ajustar as políticas, se necessário.

Thin Provisioning

O provisionamento dinâmico (thin provisioning) para um banco de dados Oracle no ASA r2 requer um planejamento cuidadoso, pois envolve a configuração de mais espaço lógico do que o fisicamente disponível. Quando implementado corretamente, o provisionamento dinâmico proporciona economias de custos significativas e melhor gerenciamento.

O provisionamento dinâmico (thin provisioning) é parte integrante do ASA r2 e está intimamente relacionado às tecnologias de eficiência do ONTAP , pois ambos permitem o armazenamento de mais dados lógicos do que a capacidade física do sistema. Os sistemas ASA r2 são exclusivamente SAN e o provisionamento dinâmico se aplica a unidades de armazenamento e LUNs dentro das Zonas de Disponibilidade de Armazenamento (SAZ).

As unidades de armazenamento ASA r2 são provisionadas dinamicamente por padrão.

Quase toda utilização de snapshots envolve provisionamento dinâmico. Por exemplo, um banco de dados típico de 10 TiB com snapshots dos últimos 30 dias pode aparecer como 310 TiB de dados lógicos, mas apenas 12 TiB a 15 TiB de espaço físico são consumidos, pois os snapshots armazenam somente os blocos alterados.

Da mesma forma, a clonagem é outra forma de provisionamento dinâmico. Um ambiente de desenvolvimento com 40 clones de um banco de dados de 80 TiB exigiria 3,2 PiB se totalmente implementado, mas na prática consome muito menos, pois apenas as alterações são armazenadas.

Gerenciamento de espaço

É preciso ter cuidado com o provisionamento dinâmico em um ambiente de aplicação, pois as taxas de alteração de dados podem aumentar inesperadamente. Por exemplo, o consumo de espaço devido a snapshots pode crescer rapidamente se as tabelas do banco de dados forem reindexadas ou se patches em larga escala forem aplicados às máquinas virtuais VMware. Um backup mal posicionado pode gravar uma grande quantidade de dados em um curto período de tempo. Por fim, pode ser difícil recuperar algumas aplicações se um LUN ficar sem espaço livre inesperadamente.

No ASA r2, esses riscos são mitigados por meio de **provisionamento dinâmico**, **monitoramento proativo** e **políticas de redimensionamento de LUN**, em vez de recursos do ONTAP como crescimento automático de volume ou exclusão automática de snapshots. Os administradores devem:

- Ativar provisionamento dinâmico em LUNs (`space-reserve disabled`) - esta é a configuração padrão no ASA r2
- Monitore a capacidade usando alertas do Gerenciador de Sistemas ou automação baseada em API.
- Use o redimensionamento de LUN agendado ou programado para acomodar o crescimento.
- Configure a reserva de snapshots e a exclusão automática de snapshots através do Gerenciador de Sistemas (GUI).

O planejamento cuidadoso dos limites de espaço e dos scripts de automação é essencial, pois o ASA r2 não oferece suporte ao crescimento automático de volumes nem à exclusão de snapshots controlada pela linha de comando.

O ASA r2 não utiliza configurações de reserva fracionária porque é uma arquitetura exclusiva de SAN que abstrai as opções de volume baseadas em WAFL. Em vez disso, a eficiência de espaço e a proteção contra sobrescrita são gerenciadas no nível do LUN. Por exemplo, se você tiver um LUN de 250 GiB provisionado a partir de uma unidade de armazenamento, os snapshots consomem espaço com base nas alterações reais de bloco, em vez de reservar uma quantidade igual de espaço antecipadamente. Isso elimina a necessidade de grandes reservas estáticas, que eram comuns em ambientes ONTAP tradicionais que utilizavam reserva fracionária.

> Caso seja necessária proteção garantida contra sobrescrita e o monitoramento não seja viável, os administradores devem provisionar capacidade suficiente na unidade de armazenamento e configurar a reserva de snapshots adequadamente. No entanto, o design do ASA r2 torna a reserva fracionária desnecessária para a maioria das cargas de trabalho.

Compactação e deduplicação

A compressão e a deduplicação no ASA r2 são tecnologias de eficiência de espaço, não mecanismos tradicionais de provisionamento dinâmico. Essas funcionalidades reduzem a necessidade de armazenamento físico, eliminando dados redundantes e comprimindo blocos, permitindo o armazenamento de mais dados lógicos do que a capacidade bruta permitiria de outra forma.

Por exemplo, um conjunto de dados de 50 TiB pode ser compactado para 30 TiB, economizando 20 TiB de espaço físico. Do ponto de vista da aplicação, ainda existem 50 TiB de dados, embora ocupem apenas 30 TiB em disco.

> A compressibilidade de um conjunto de dados pode mudar ao longo do tempo, o que pode aumentar o consumo de espaço físico. Portanto, a compressão e a desduplicação devem ser gerenciadas proativamente por meio de monitoramento e planejamento de capacidade.

Espaço livre e alocação de espaço LVM

O provisionamento dinâmico em ambientes ASA r2 pode perder eficiência ao longo do tempo se os blocos excluídos não forem recuperados. A menos que o espaço seja liberado usando TRIM/UNMAP ou sobrescrito com zeros (via ASMRU - Utilitário Automático de Gerenciamento e Recuperação de Espaço), os dados excluídos continuam a consumir capacidade física. Em muitos ambientes de banco de dados Oracle, o provisionamento dinâmico oferece benefícios limitados, pois os arquivos de dados normalmente são pré-alocados em seu tamanho total durante a criação.

Um planejamento cuidadoso da configuração do LVM pode melhorar a eficiência e minimizar a necessidade de provisionamento de armazenamento e redimensionamento de LUNs. Quando um LVM como o Veritas VxVM ou o Oracle ASM é utilizado, os LUNs subjacentes são divididos em extensões que são usadas somente quando necessário. Por exemplo, se um conjunto de dados começar com 2 TiB de tamanho, mas puder crescer para 10 TiB ao longo do tempo, esse conjunto de dados poderá ser colocado em 10 TiB de LUNs com provisionamento dinâmico, organizadas em um grupo de discos LVM. Ocuparia apenas 2 TiB de espaço no momento da criação e só exigiria espaço adicional à medida que extensões fossem alocadas para acomodar o crescimento dos dados. Este processo é seguro desde que o espaço seja monitorado.

Failover do ONTAP

É necessário compreender as funções de transferência de armazenamento para garantir que as operações do banco de dados Oracle não sejam interrompidas durante essas operações. Além disso, os argumentos utilizados em operações de aquisição podem afetar a integridade dos dados se forem usados incorretamente.

Em condições normais, as gravações recebidas por um determinado controlador são espelhadas de forma síncrona em seu parceiro de alta disponibilidade (HA). Em um ambiente ASA r2 com SnapMirror Active Sync (SM-as), as gravações também são espelhadas para um controlador remoto no site secundário. Até que uma gravação seja armazenada em mídia não volátil em todos os locais, ela não será confirmada para o aplicativo host.

O meio que armazena os dados de gravação é chamado de memória não volátil (NVMEM). Às vezes é chamada de memória de acesso aleatório não volátil (NVRAM) e pode ser considerada um diário de escrita em vez de um cache. Durante o funcionamento normal, os dados da NVMEM não são lidos; ela é usada apenas para proteger os dados em caso de falha de software ou hardware. Quando os dados são gravados nas unidades, eles são transferidos da RAM do sistema, e não da NVMEM.

Durante uma operação de tomada de controle, um nó em um par de alta disponibilidade assume as operações de seu parceiro. No ASA r2, a troca de recursos não é aplicável porque o MetroCluster não é suportado; em vez disso, o SnapMirror Active Sync fornece redundância em nível de site. As operações de transferência de armazenamento durante a manutenção de rotina devem ser transparentes, exceto por uma breve pausa nas operações enquanto os caminhos de rede são alterados. A criação de redes pode ser complexa e erros são fáceis de cometer, por isso a NetApp recomenda enfaticamente testar minuciosamente as operações de takeover antes de colocar um sistema de armazenamento em produção. Essa é a única maneira de garantir que todos os caminhos de rede estejam configurados corretamente. Em um ambiente SAN, verifique o status do caminho usando o comando `sanlun lun show -p` ou as ferramentas nativas de multipathing do sistema operacional para garantir que todos os caminhos esperados estejam disponíveis. Os sistemas ASA r2 fornecem todos os caminhos ativos otimizados para LUNs, e os clientes que usam namespaces NVMe devem depender de ferramentas nativas do sistema operacional, pois os caminhos NVMe não são cobertos pelo sanlun.

É preciso ter cautela ao decretar uma tomada de posse forçada. Forçar uma alteração na configuração de armazenamento significa que o estado do controlador que detém as unidades é desconsiderado e o nó alternativo assume o controle das unidades à força. A execução incorreta de uma operação de takeover pode resultar em perda ou corrupção de dados, pois um takeover forçado pode descartar o conteúdo da NVMEM. Após a conclusão da aquisição, a perda desses dados significa que os dados armazenados nos discos podem retornar a um estado ligeiramente anterior do ponto de vista do banco de dados.

Uma tomada de controle forçada com um par HA normal raramente deve ser necessária. Em praticamente todos os cenários de falha, um nó é desligado e informa o parceiro para que ocorra uma transição automática. Existem alguns casos extremos, como uma falha sequencial em que a interconexão entre os nós é perdida e, em seguida, um dos controladores falha, situação em que é necessária uma intervenção forçada. Nessa situação, o espelhamento entre os nós é perdido antes da falha do controlador, o que significa que o controlador sobrevivente não possui mais uma cópia das gravações em andamento. A tomada de controle precisa então ser forçada, o que significa que dados podem ser perdidos.

A NetApp recomenda tomar as seguintes precauções:

- Tenha muito cuidado para não forçar uma aquisição acidentalmente. Normalmente, forçar a alteração não é necessário e pode causar perda de dados.
- Caso seja necessária uma tomada de controle forçada, certifique-se de que os aplicativos estejam fechados, todos os sistemas de arquivos estejam desmontados e os grupos de volumes do gerenciador de volumes lógicos (LVM) estejam desativados (variegate). Os grupos de discos ASM devem ser desmontados.
- Em caso de falha no nível do site ao usar SM-as, o failover automático não planejado assistido pelo ONTAP Mediator será iniciado no cluster sobrevivente, resultando em uma breve pausa de E/S e, em seguida, as transições de banco de dados continuarão a partir do cluster sobrevivente. Para obter mais informações, consulte o "Sincronização ativa do SnapMirror em sistemas ASA r2" Para obter instruções detalhadas de configuração.

Configuração de banco de dados com sistemas AFF/ FAS

Tamanhos de blocos

O ONTAP usa internamente um tamanho de bloco variável, o que significa que os bancos de dados Oracle podem ser configurados com qualquer tamanho de bloco desejado. No entanto, os tamanhos de bloco de sistema de arquivos podem afetar o desempenho e, em alguns casos, um tamanho maior de bloco de refazer pode melhorar o desempenho.

Tamanhos de blocos de arquivos de dados

Alguns sistemas operacionais oferecem uma escolha de tamanhos de bloco de sistema de arquivos. Para sistemas de arquivos que suportam datafiles Oracle, o tamanho do bloco deve ser 8KB quando a compactação é usada. Quando a compressão não é necessária, um tamanho de bloco de 8KB ou 4KB pode ser usado.

Se um arquivo de dados for colocado em um sistema de arquivos com um bloco de 512 bytes, arquivos desalinhados são possíveis. O LUN e o sistema de arquivos podem estar alinhados adequadamente com base nas recomendações do NetApp, mas a e/S do arquivo estaria desalinhada. Tal desalinhamento causaria graves problemas de desempenho.

Os sistemas de arquivos que suportam os logs de refazer devem usar um tamanho de bloco que seja um múltiplo do tamanho do bloco de refazer. Isso geralmente requer que o sistema de arquivos de log refazer e o log refazer usem um tamanho de bloco de 512 bytes.

Refazer tamanhos de bloco

Com taxas de recozimento muito altas, é possível que 4KB tamanhos de bloco tenham um melhor desempenho, pois altas taxas de recozimento permitem que a e/S seja executada em operações cada vez menores e mais eficientes. Se as taxas de refazer forem maiores que 50Mbps, considere testar um tamanho de bloco 4KB.

Alguns problemas de clientes foram identificados com bancos de dados usando logs de refazer com um tamanho de bloco de 512 bytes em um sistema de arquivos com um tamanho de bloco de 4KB e muitas transações muito pequenas. A sobrecarga envolvida na aplicação de várias alterações de 512 bytes a um único bloco de sistema de arquivos 4KB levou a problemas de desempenho que foram resolvidos alterando o

sistema de arquivos para usar um tamanho de bloco de 512 bytes.

A NetApp recomenda que você não altere o tamanho do bloco refazer, a menos que seja aconselhado por uma organização de suporte ao cliente ou serviços profissionais relevante ou que a alteração seja baseada na documentação oficial do produto.

db_file_multiblock_read_count

O `db_file_multiblock_read_count` parâmetro controla o número máximo de blocos de banco de dados Oracle que o Oracle lê como uma única operação durante a e/S sequencial

Este parâmetro não afeta, no entanto, o número de blocos que o Oracle lê durante todas e quaisquer operações de leitura, nem afeta e/S aleatórias Apenas o tamanho de bloco de e/S sequenciais é afetado.

A Oracle recomenda que o usuário deixe esse parâmetro desconfigurado. Isso permite que o software de banco de dados defina automaticamente o valor ideal. Isso geralmente significa que esse parâmetro é definido para um valor que produz um tamanho de e/S de 1MB. Por exemplo, uma leitura 1MB de 8KB blocos exigiria que 128 blocos fossem lidos, e o valor padrão para este parâmetro seria, portanto, 128.

A maioria dos problemas de desempenho do banco de dados observados pelo NetApp nas instalações dos clientes envolve uma configuração incorreta para este parâmetro. Houve razões válidas para alterar esse valor com as versões 8 e 9 do Oracle. Como resultado, o parâmetro pode estar inconscientemente presente em `init.ora` arquivos porque o banco de dados foi atualizado para o Oracle 10 e posterior. Uma configuração legada de 8 ou 16, comparada a um valor padrão de 128, danifica significativamente o desempenho de e/S sequenciais.

NetApp recomenda a configuração `db_file_multiblock_read_count` do parâmetro não deve estar presente no `init.ora` arquivo. O NetApp nunca encontrou uma situação em que a alteração desse parâmetro melhorou o desempenho, mas há muitos casos em que causou danos claros ao throughput sequencial de e/S.

sistema de arquivos_options

O parâmetro de inicialização do Oracle `filesystemio_options` controla o uso de e/S assíncrona e direta

Ao contrário da crença comum, a e/S assíncrona e direta não são mutuamente exclusivas. A NetApp observou que esse parâmetro é frequentemente mal configurado em ambientes de clientes, e essa configuração incorreta é diretamente responsável por muitos problemas de desempenho.

E/S assíncrona significa que as operações de e/S Oracle podem ser paralelizadas. Antes da disponibilidade de e/S assíncrona em vários sistemas operacionais, os usuários configuraram vários processos dbwriter e alteraram a configuração do processo do servidor. Com e/S assíncrona, o próprio sistema operacional executa e/S em nome do software de banco de dados de maneira altamente eficiente e paralela. Esse processo não coloca os dados em risco, e operações críticas, como o Oracle refazer o log, ainda são executadas de forma síncrona.

A e/S direta ignora o cache do buffer do sistema operacional. E/S em um sistema UNIX normalmente flui através do cache de buffer do sistema operacional. Isso é útil para aplicativos que não mantêm um cache interno, mas o Oracle tem seu próprio cache de buffer dentro do SGA. Em quase todos os casos, é melhor habilitar e/S direta e alocar RAM do servidor para o SGA em vez de confiar no cache de buffer do sistema

operacional. O Oracle SGA usa a memória de forma mais eficiente. Além disso, quando a e/S flui pelo buffer do sistema operacional, ela está sujeita a processamento adicional, o que aumenta as latências. As latências aumentadas são especialmente percetíveis com e/S de gravação intensa quando a baixa latência é um requisito essencial.

As opções para `filesystemio_options` são:

- **assíncrono.** A Oracle envia solicitações de e/S para o sistema operacional para processamento. Esse processo permite que a Oracle execute outro trabalho em vez de esperar pela conclusão de e/S e, assim, aumenta a paralelização de e/S.
- **directio.** A Oracle executa e/S diretamente contra arquivos físicos em vez de rotear e/S pelo cache do sistema operacional do host.
- **nenhum.** A Oracle usa e/S síncrona e em buffer Nesta configuração, a escolha entre processos de servidor compartilhado e dedicado e o número de dbwriters são mais importantes.
- * setall.* A Oracle usa e/S assíncrona e direta Em quase todos os casos, o uso de `setall` é ideal.

O `filesystemio_options` parâmetro não tem efeito em ambientes DNFS e ASM. O uso de DNFS ou ASM resulta automaticamente no uso de e/S assíncrona e direta

Alguns clientes encontraram problemas assíncronos de I/o no passado, especialmente com versões anteriores do Red Hat Enterprise Linux 4 (RHEL4). Alguns conselhos desatualizados na internet ainda sugerem evitar IO assíncrono por causa de informações desatualizadas. E/S assíncrona é estável em todos os sistemas operacionais atuais. Não há razão para desativá-lo, ausente um bug conhecido com o sistema operacional.

Se um banco de dados tiver usado e/S em buffer, um switch para direcionar e/S também pode garantir uma alteração no tamanho do SGA. A desativação da e/S armazenada em buffer elimina o benefício de desempenho que o cache do sistema operacional do host fornece para o banco de dados. Adicionar RAM de volta ao SGA repara este problema. O resultado líquido deve ser uma melhoria no desempenho de e/S.

Embora seja quase sempre melhor usar RAM para o Oracle SGA do que para cache de buffer do SO, pode ser impossível determinar o melhor valor. Por exemplo, pode ser preferível usar e/S em buffer com tamanhos SGA muito pequenos em um servidor de banco de dados com muitas instâncias Oracle ativas intermitentemente. Este arranjo permite o uso flexível da RAM livre restante no sistema operacional por todas as instâncias de banco de dados em execução. Esta é uma situação altamente incomum, mas tem sido observada em alguns sites de clientes.

A NetApp recomenda a configuração `filesystemio_options` para `setall`, mas esteja ciente de que, em algumas circunstâncias, a perda do cache de buffer do host pode exigir um aumento no SGA Oracle.

Tempos limite do RAC

O Oracle RAC é um produto exclusivo com vários tipos de processos internos de heartbeat que monitoram a integridade do cluster.

As informações na "número de identificação" seção incluem informações críticas para ambientes Oracle RAC que usam armazenamento em rede e, em muitos casos, as configurações padrão do Oracle RAC precisarão ser alteradas para garantir que o cluster RAC sobreviva às alterações do caminho da rede e às operações de failover/switchover de armazenamento.

disktimeout

O parâmetro RAC principal relacionado ao armazenamento é `disktimeout`. Este parâmetro controla o limite dentro do qual a e/S do arquivo de votação deve ser concluída. Se o `disktimeout` parâmetro for excedido, o nó RAC será despejado do cluster. O padrão para este parâmetro é 200. Este valor deve ser suficiente para os procedimentos normais de aquisição de armazenamento e de giveback.

A NetApp recomenda fortemente testar cuidadosamente as configurações do RAC antes de colocá-las em produção, pois muitos fatores afetam uma aquisição ou a giveback. Além do tempo necessário para a conclusão do failover de armazenamento, também é necessário tempo adicional para que as alterações do protocolo de controle de agregação de link (LACP) se propaguem. Além disso, o software de multipathing SAN deve detetar um tempo limite de e/S e tentar novamente em um caminho alternativo. Se um banco de dados estiver extremamente ativo, uma grande quantidade de e/S deve ser enfileirada e tentada novamente antes de o disco de votação ser processado.

Se não for possível executar uma aquisição de armazenamento real ou giveback, o efeito poderá ser simulado com testes de pull de cabo no servidor de banco de dados.

A NetApp recomenda o seguinte:

- Deixando o `disktimeout` parâmetro no valor padrão de 200.
- Sempre teste uma configuração RAC completamente.

número de identificação

O `misscount` parâmetro normalmente afeta apenas o batimento cardíaco da rede entre nós RAC. O padrão é 30 segundos. Se os binários de grade estiverem em um storage array ou a unidade de inicialização do sistema operacional não for local, esse parâmetro pode se tornar importante. Isso inclui hosts com unidades de inicialização localizadas em uma SAN FC, sistemas operacionais iniciados por NFS e unidades de inicialização localizados em datastores de virtualização, como um arquivo VMDK.

Se o acesso a uma unidade de inicialização for interrompido por uma aquisição de armazenamento ou giveback, é possível que a localização binária da grade ou todo o sistema operacional fique temporariamente suspenso. O tempo necessário para o ONTAP concluir a operação de storage e para o sistema operacional alterar caminhos e retomar e/S pode exceder o `misscount` limite. Como resultado, um nó é despejado imediatamente após a conetividade com o LUN de inicialização ou binários de grade ser restaurada. Na maioria dos casos, o despejo e a reinicialização subsequente ocorrem sem mensagens de Registro para indicar o motivo da reinicialização. Nem todas as configurações são afetadas, portanto, teste qualquer host baseado em SAN-boot, NFS-boot ou datastore em um ambiente RAC para que o RAC permaneça estável se a comunicação com a unidade de inicialização for interrompida.

No caso de unidades de inicialização não locais ou binários de hospedagem de sistemas de arquivos não locais `grid`, o `misscount` precisará ser alterado para corresponder `disktimeout`ao . Se esse parâmetro for alterado, realize testes adicionais para identificar também quaisquer efeitos no comportamento do RAC, como o tempo de failover do nó.

A NetApp recomenda o seguinte:

- Deixe o `misscount` parâmetro no valor padrão de 30, a menos que uma das seguintes condições se aplique:
  - `grid` Os binários estão localizados em uma unidade conetada à rede, incluindo unidades baseadas em NFS, iSCSI, FC e datastore.
  - O sistema operacional é inicializado pela SAN.
- Nesses casos, avalie o efeito das interrupções da rede que afetam o acesso ao sistema operacional ou `GRID_HOME` aos sistemas de arquivos. Em alguns casos, tais interrupções fazem com que os daemons do Oracle RAC parem, o que pode levar a um `misscount` tempo limite e despejo baseado em -. O tempo limite padrão é 27 segundos, que é o valor de `misscount` menos `reboottime`. Nesses casos, aumente `misscount` para 200 para corresponder `disktimeout`.

Configuração de banco de dados com sistemas ASA r2

Tamanhos de blocos

Internamente, o ONTAP utiliza um tamanho de bloco variável, o que significa que os bancos de dados Oracle podem ser configurados com qualquer tamanho de bloco desejado. No entanto, o tamanho dos blocos do sistema de arquivos pode afetar o desempenho e, em alguns casos, um tamanho maior para o bloco de refazer pode melhorar o desempenho.

O ASA r2 não introduz nenhuma alteração nas recomendações de tamanho de bloco do Oracle em comparação com os sistemas AFF/ FAS . O comportamento do ONTAP permanece consistente em todas as plataformas.

Tamanhos de blocos de arquivos de dados

Alguns sistemas operacionais oferecem uma escolha de tamanhos de bloco de sistema de arquivos. Para sistemas de arquivos que suportam datafiles Oracle, o tamanho do bloco deve ser 8KB quando a compactação é usada. Quando a compressão não é necessária, um tamanho de bloco de 8KB ou 4KB pode ser usado.

Se um arquivo de dados for colocado em um sistema de arquivos com um bloco de 512 bytes, arquivos desalinhados são possíveis. O LUN e o sistema de arquivos podem estar alinhados adequadamente com base nas recomendações do NetApp, mas a e/S do arquivo estaria desalinhada. Tal desalinhamento causaria graves problemas de desempenho.

Refazer tamanhos de bloco

Os sistemas de arquivos que suportam os logs de refazer devem usar um tamanho de bloco que seja um múltiplo do tamanho do bloco de refazer. Isso geralmente requer que o sistema de arquivos de log refazer e o log refazer usem um tamanho de bloco de 512 bytes.

Com taxas de recozimento muito altas, é possível que 4KB tamanhos de bloco tenham um melhor desempenho, pois altas taxas de recozimento permitem que a e/S seja executada em operações cada vez menores e mais eficientes. Se as taxas de refazer forem maiores que 50Mbps, considere testar um tamanho de bloco 4KB.

Alguns problemas de clientes foram identificados com bancos de dados usando logs de refazer com um tamanho de bloco de 512 bytes em um sistema de arquivos com um tamanho de bloco de 4KB e muitas transações muito pequenas. A sobrecarga envolvida na aplicação de várias alterações de 512 bytes a um único bloco de sistema de arquivos 4KB levou a problemas de desempenho que foram resolvidos alterando o sistema de arquivos para usar um tamanho de bloco de 512 bytes.

A NetApp recomenda que você não altere o tamanho do bloco refazer, a menos que seja aconselhado por uma organização de suporte ao cliente ou serviços profissionais relevante ou que a alteração seja baseada na documentação oficial do produto.

db_file_multiblock_read_count

O `db_file_multiblock_read_count` parâmetro controla o número máximo de blocos de banco de dados Oracle que o Oracle lê como uma única operação durante a e/S sequencial

Não há alterações nas recomendações em comparação com os sistemas AFF/ FAS . O comportamento do ONTAP e as melhores práticas da Oracle permanecem idênticos nas plataformas ASA r2, AFF e FAS .

Este parâmetro não afeta, no entanto, o número de blocos que o Oracle lê durante todas e quaisquer operações de leitura, nem afeta e/S aleatórias Apenas o tamanho de bloco de e/S sequenciais é afetado.

A Oracle recomenda que o usuário deixe esse parâmetro desconfigurado. Isso permite que o software de banco de dados defina automaticamente o valor ideal. Isso geralmente significa que esse parâmetro é definido para um valor que produz um tamanho de e/S de 1MB. Por exemplo, uma leitura 1MB de 8KB blocos exigiria que 128 blocos fossem lidos, e o valor padrão para este parâmetro seria, portanto, 128.

A maioria dos problemas de desempenho do banco de dados observados pelo NetApp nas instalações dos clientes envolve uma configuração incorreta para este parâmetro. Houve razões válidas para alterar esse valor com as versões 8 e 9 do Oracle. Como resultado, o parâmetro pode estar inconscientemente presente em `init.ora` arquivos porque o banco de dados foi atualizado para o Oracle 10 e posterior. Uma configuração legada de 8 ou 16, comparada a um valor padrão de 128, danifica significativamente o desempenho de e/S sequenciais.

NetApp recomenda a configuração `db_file_multiblock_read_count` do parâmetro não deve estar presente no `init.ora` arquivo. O NetApp nunca encontrou uma situação em que a alteração desse parâmetro melhorou o desempenho, mas há muitos casos em que causou danos claros ao throughput sequencial de e/S.

sistema de arquivos_options

O parâmetro de inicialização do Oracle `filesystemio_options` controla o uso de e/S assíncrona e direta

O comportamento e as recomendações para filesystemio_options no ASA r2 são idênticos aos dos sistemas AFF/ FAS , pois o parâmetro é específico da Oracle e não depende da plataforma de armazenamento. O ASA r2 usa o ONTAP , assim como o AFF/ FAS, portanto, as mesmas boas práticas se aplicam.

Ao contrário da crença comum, a e/S assíncrona e direta não são mutuamente exclusivas. A NetApp observou que esse parâmetro é frequentemente mal configurado em ambientes de clientes, e essa configuração incorreta é diretamente responsável por muitos problemas de desempenho.

E/S assíncrona significa que as operações de e/S Oracle podem ser paralelizadas. Antes da disponibilidade de e/S assíncrona em vários sistemas operacionais, os usuários configuraram vários processos dbwriter e alteraram a configuração do processo do servidor. Com e/S assíncrona, o próprio sistema operacional executa e/S em nome do software de banco de dados de maneira altamente eficiente e paralela. Esse processo não coloca os dados em risco, e operações críticas, como o Oracle refazer o log, ainda são executadas de forma síncrona.

A e/S direta ignora o cache do buffer do sistema operacional. E/S em um sistema UNIX normalmente flui através do cache de buffer do sistema operacional. Isso é útil para aplicativos que não mantêm um cache interno, mas o Oracle tem seu próprio cache de buffer dentro do SGA. Em quase todos os casos, é melhor habilitar e/S direta e alocar RAM do servidor para o SGA em vez de confiar no cache de buffer do sistema operacional. O Oracle SGA usa a memória de forma mais eficiente. Além disso, quando a e/S flui pelo buffer do sistema operacional, ela está sujeita a processamento adicional, o que aumenta as latências. As latências aumentadas são especialmente percetíveis com e/S de gravação intensa quando a baixa latência é um requisito essencial.

As opções para `filesystemio_options` são:

- **assíncrono.** A Oracle envia solicitações de e/S para o sistema operacional para processamento. Esse processo permite que a Oracle execute outro trabalho em vez de esperar pela conclusão de e/S e, assim, aumenta a paralelização de e/S.
- **directio.** A Oracle executa e/S diretamente contra arquivos físicos em vez de rotear e/S pelo cache do sistema operacional do host.
- **nenhum.** A Oracle usa e/S síncrona e em buffer Nesta configuração, a escolha entre processos de servidor compartilhado e dedicado e o número de dbwriters são mais importantes.
- * setall.* A Oracle usa e/S assíncrona e direta Em quase todos os casos, o uso de `setall` é ideal.

> Em ambientes ASM, o Oracle usa automaticamente E/S direta e E/S assíncrona para discos gerenciados pelo ASM. `filesystemio_options` Não tem efeito sobre os grupos de discos ASM. Para implantações que não utilizam ASM (por exemplo, sistemas de arquivos em LUNs SAN), defina: `filesystemio_options = setall`. Isso possibilita tanto E/S assíncrona quanto direta para um desempenho ideal.

Alguns sistemas operacionais mais antigos apresentavam problemas com E/S assíncrona, o que levou a recomendações desatualizadas sugerindo que ela fosse evitada. No entanto, a E/S assíncrona é estável e totalmente suportada em todos os sistemas operacionais atuais. Não há motivo para desativá-lo, a menos que seja identificado um bug específico do sistema operacional.

Se um banco de dados tiver usado e/S em buffer, um switch para direcionar e/S também pode garantir uma alteração no tamanho do SGA. A desativação da e/S armazenada em buffer elimina o benefício de desempenho que o cache do sistema operacional do host fornece para o banco de dados. Adicionar RAM de volta ao SGA repara este problema. O resultado líquido deve ser uma melhoria no desempenho de e/S.

Embora seja quase sempre melhor usar RAM para o Oracle SGA do que para cache de buffer do SO, pode ser impossível determinar o melhor valor. Por exemplo, pode ser preferível usar e/S em buffer com tamanhos SGA muito pequenos em um servidor de banco de dados com muitas instâncias Oracle ativas intermitentemente. Este arranjo permite o uso flexível da RAM livre restante no sistema operacional por todas as instâncias de banco de dados em execução. Esta é uma situação altamente incomum, mas tem sido observada em alguns sites de clientes.

* A NetApp recomenda* a configuração `filesystemio_options` para `setall`Mas esteja ciente de que, em algumas circunstâncias, a perda do cache de buffer do host pode exigir um aumento na SGA do Oracle. Os sistemas ASA r2 são otimizados para cargas de trabalho SAN com baixa latência, portanto, o uso do setall está perfeitamente alinhado com o design do ASA para implementações Oracle de alto desempenho.

Tempos limite do RAC

O Oracle RAC é um produto exclusivo com vários tipos de processos internos de heartbeat que monitoram a integridade do cluster.

Os sistemas ASA r2 usam ONTAP da mesma forma que o AFF/ FAS, portanto, os mesmos princípios se aplicam aos parâmetros de tempo limite do Oracle RAC. Não há alterações específicas do ASA nas recomendações de tempo limite de disco ou contagem de erros. No entanto, o ASA r2 é otimizado para cargas de trabalho SAN e failover de baixa latência, o que torna essas práticas recomendadas ainda mais críticas.

As informações em "número de identificação" Esta seção inclui informações críticas para ambientes Oracle RAC que utilizam armazenamento em rede e, em muitos casos, as configurações padrão do Oracle RAC precisarão ser alteradas para garantir que o cluster RAC sobreviva a alterações no caminho da rede e operações de failover de armazenamento.

disktimeout

O parâmetro RAC principal relacionado ao armazenamento é `disktimeout`. Este parâmetro controla o limite dentro do qual a e/S do arquivo de votação deve ser concluída. Se o `disktimeout` parâmetro for excedido, o nó RAC será despejado do cluster. O padrão para este parâmetro é 200. Este valor deve ser suficiente para os procedimentos normais de aquisição de armazenamento e de giveback.

A NetApp recomenda fortemente testar cuidadosamente as configurações do RAC antes de colocá-las em produção, pois muitos fatores afetam uma aquisição ou a giveback. Além do tempo necessário para a conclusão do failover de armazenamento, também é necessário tempo adicional para que as alterações do protocolo de controle de agregação de link (LACP) se propaguem. Além disso, o software de multipathing SAN deve detetar um tempo limite de e/S e tentar novamente em um caminho alternativo. Se um banco de dados estiver extremamente ativo, uma grande quantidade de e/S deve ser enfileirada e tentada novamente antes de o disco de votação ser processado.

Se não for possível executar uma aquisição de armazenamento real ou giveback, o efeito poderá ser simulado com testes de pull de cabo no servidor de banco de dados.

A NetApp recomenda o seguinte:

- Deixando o `disktimeout` parâmetro no valor padrão de 200.
- Sempre teste uma configuração RAC completamente.

número de identificação

O `misscount` parâmetro normalmente afeta apenas o batimento cardíaco da rede entre nós RAC. O padrão é 30 segundos. Se os binários de grade estiverem em um storage array ou a unidade de inicialização do sistema operacional não for local, esse parâmetro pode se tornar importante. Isso inclui hosts com unidades de inicialização localizadas em uma SAN FC, sistemas operacionais iniciados por NFS e unidades de

inicialização localizados em datastores de virtualização, como um arquivo VMDK.

Se o acesso a uma unidade de inicialização for interrompido por uma aquisição de armazenamento ou giveback, é possível que a localização binária da grade ou todo o sistema operacional fique temporariamente suspenso. O tempo necessário para o ONTAP concluir a operação de storage e para o sistema operacional alterar caminhos e retomar e/S pode exceder o `misscount` limite. Como resultado, um nó é despejado imediatamente após a conetividade com o LUN de inicialização ou binários de grade ser restaurada. Na maioria dos casos, o despejo e a reinicialização subsequente ocorrem sem mensagens de Registro para indicar o motivo da reinicialização. Nem todas as configurações são afetadas, portanto, teste qualquer host baseado em SAN-boot, NFS-boot ou datastore em um ambiente RAC para que o RAC permaneça estável se a comunicação com a unidade de inicialização for interrompida.

No caso de unidades de inicialização não locais ou binários de hospedagem de sistemas de arquivos não locais `grid`, o `misscount` precisará ser alterado para corresponder `disktimeout`ao . Se esse parâmetro for alterado, realize testes adicionais para identificar também quaisquer efeitos no comportamento do RAC, como o tempo de failover do nó.

A NetApp recomenda o seguinte:

- Deixe o `misscount` parâmetro no valor padrão de 30, a menos que uma das seguintes condições se aplique:
  - `grid` Os arquivos binários estão localizados em uma unidade conectada à rede, incluindo unidades iSCSI, FC e unidades baseadas em armazenamento de dados.
  - O sistema operacional é inicializado pela SAN.
- Nesses casos, avalie o efeito das interrupções da rede que afetam o acesso ao sistema operacional ou `GRID_HOME` aos sistemas de arquivos. Em alguns casos, tais interrupções fazem com que os daemons do Oracle RAC parem, o que pode levar a um `misscount` tempo limite e despejo baseado em -. O tempo limite padrão é 27 segundos, que é o valor de `misscount` menos `reboottime`. Nesses casos, aumente `misscount` para 200 para corresponder `disktimeout`.

- O design otimizado para SAN do ASA r2 reduz a latência de failover, mas os tempos limite ainda precisam ser ajustados para inicialização em rede ou binários de grid.
- Para configurações RAC extensas ou ativas-ativas (por exemplo, sincronização ativa SnapMirror), o ajuste de tempo limite continua sendo essencial para arquiteturas de RPO zero.

Configuração de host com sistemas AFF/ FAS

AIX

Tópicos de configuração para banco de dados Oracle no IBM AIX com ONTAP.

E/S simultânea

Alcançar o desempenho ideal no IBM AIX requer o uso de e/S simultâneas Sem I/o concorrente, as limitações de desempenho provavelmente são porque o AIX executa e/S atômica serializada, o que incorre em sobrecarga significativa.

Originalmente, o NetApp recomendou o uso da `cio` opção de montagem para forçar o uso de e/S concorrente

no sistema de arquivos, mas esse processo teve desvantagens e não é mais necessário. Desde a introdução do AIX 5,2 e do Oracle 10gR1, o Oracle no AIX pode abrir arquivos individuais para IO simultâneo, em vez de forçar e/S simultânea em todo o sistema de arquivos.

O melhor método para ativar e/S concorrente é definir o `init.ora` parâmetro `filesystemio_options` como `setall`. Isso permite que a Oracle abra arquivos específicos para uso com e/S concorrente

Usar `cio` como opção de montagem força o uso de e/S concorrente, o que pode ter consequências negativas. Por exemplo, forçar e/S concorrente desativa o readahead em sistemas de arquivos, o que pode danificar o desempenho de e/S que ocorre fora do software de banco de dados Oracle, como copiar arquivos e executar backups de fita. Além disso, produtos como Oracle GoldenGate e SAP BR*Tools não são compatíveis com o uso da `cio` opção de montagem com certas versões do Oracle.

A NetApp recomenda o seguinte:

- Não use a `cio` opção de montagem no nível do sistema de arquivos. Em vez disso, ative a e/S concorrente através do uso `filesystemio_options=setall` do .
- Utilize apenas a `cio` opção de montagem se não for possível definir `filesystemio_options=setall`.

Opções de montagem AIX NFS

A tabela a seguir lista as opções de montagem NFS AIX para bancos de dados de instância única Oracle.

Tipo de ficheiro	Opções de montagem
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Registros do Redo ControlFiles Datafiles	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
`ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, intr`

A tabela a seguir lista as opções de montagem NFS AIX para RAC.

Tipo de ficheiro	Opções de montagem
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Registros do Redo ControlFiles Datafiles	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac`
`CRS/Voting`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac`
Dedicado `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`

Tipo de ficheiro	Opções de montagem
Compartilhado `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

A principal diferença entre as opções de montagem de instância única e RAC é a adição `noac` das opções de montagem. Essa adição tem o efeito de desabilitar o cache do sistema operacional do host, que permite que todas as instâncias do cluster RAC tenham uma visão consistente do estado dos dados.

Embora o uso da `cio` opção de montagem e do `init.ora` parâmetro `filesystemio_options=setall` tenha o mesmo efeito de desabilitar o cache do host, ainda é necessário usar `noac`o . `noac` É necessário para implantações compartilhadas `ORACLE_HOME` para facilitar a consistência de arquivos, como arquivos de senha Oracle e `spfile` arquivos de parâmetros. Se cada instância em um cluster RAC tiver um dedicado `ORACLE_HOME`, esse parâmetro não será necessário.

AIX jfs/jfs2 Opções de montagem

A tabela a seguir lista as opções de montagem do AIX jfs/jfs2.

Tipo de ficheiro	Opções de montagem
ADR Home	Predefinições
Registros do Redo ControlFiles Datafiles	Predefinições
ORACLE_HOME	Predefinições

Antes de usar dispositivos AIX `hdisk` em qualquer ambiente, incluindo bancos de dados, verifique o parâmetro `queue_depth`. Este parâmetro não é a profundidade da fila HBA; em vez disso, ele se relaciona com a profundidade da fila SCSI do indivíduo `hdisk device. Depending on how the LUNs are configured, the value for `queue_depth` pode ser muito baixa para um bom desempenho. Os testes mostraram que o valor ideal é 64.

HP-UX

Tópicos de configuração para banco de dados Oracle no HP-UX com ONTAP.

Opções de montagem NFS HP-UX

A tabela a seguir lista as opções de montagem HP-UX NFS para uma única instância.

Tipo de ficheiro	Opções de montagem
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, suid`
Arquivos de controle Datafiles Redo logs	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144,forc edirectio, nointr,suid`

Tipo de ficheiro	Opções de montagem
ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid

A tabela a seguir lista as opções de montagem HP-UX NFS para RAC.

Tipo de ficheiro	Opções de montagem
ADR Home	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,noac,suid
Arquivos de controle Datafiles Redo logs	rw, bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio,suid
CRS/votação	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio,suid
Dedicado ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid
Compartilhado ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,suid

A principal diferença entre as opções de montagem de uma única instância e RAC é a adição `noac` de e `forcedirectio` às opções de montagem. Essa adição tem o efeito de desabilitar o cache do sistema operacional do host, o que permite que todas as instâncias do cluster RAC tenham uma visão consistente do estado dos dados. Embora o uso do `init.ora` parâmetro `filesystemio_options=setall` tenha o mesmo efeito de desabilitar o cache do host, ainda é necessário usar `noac` e `forcedirectio`.

O motivo `noac` necessário para implantações compartilhadas `ORACLE_HOME` é facilitar a consistência de arquivos, como arquivos de senha Oracle e spfiles. Se cada instância em um cluster RAC tiver um dedicado `ORACLE_HOME`, esse parâmetro não será necessário.

Opções de montagem HP-UX VxFS

Use as seguintes opções de montagem para sistemas de arquivos que hospedam binários Oracle:

```
delaylog,nodatainlog
```

Use as seguintes opções de montagem para sistemas de arquivos que contêm datafiles, logs de refazer, logs de arquivamento e arquivos de controle nos quais a versão do HP-UX não suporta e/S simultânea:

```
nodatainlog,mincache=direct,convosync=direct
```

Quando a e/S simultânea for suportada (VxFS 5.0.1 e posterior, ou com o ServiceGuard Storage Management Suite), utilize estas opções de montagem para sistemas de ficheiros que contenham ficheiros de dados, registos refeitos, registos de arquivo e ficheiros de controlo:

```
delaylog,cio
```

O parâmetro `db_file_multiblock_read_count` é especialmente crítico em ambientes VxFS. A Oracle recomenda que esse parâmetro permaneça desconfigurado no Oracle 10g R1i e posterior, a menos que especificamente direcionado de outra forma. O padrão com um tamanho de bloco Oracle 8KB é 128. Se o valor deste parâmetro for forçado a 16 ou menos, remova a `convosync=direct` opção de montagem porque pode danificar o desempenho sequencial de e/S. Esta etapa prejudica outros aspetos do desempenho e só deve ser tomada se o valor de `db_file_multiblock_read_count` tiver de ser alterado do valor padrão.

Linux

Tópicos de configuração específicos para o sistema operacional Linux.

Tabelas de slots TCP do Linux NFSv3

As tabelas de slot TCP são equivalentes a NFSv3 mm de profundidade de fila do adaptador de barramento do host (HBA). Essas tabelas controlam o número de operações NFS que podem ficar pendentes de uma só vez. O valor padrão é geralmente 16, o que é muito baixo para um desempenho ideal. O problema oposto ocorre em kernels Linux mais recentes, que podem aumentar automaticamente o limite da tabela de slots TCP para um nível que satura o servidor NFS com solicitações.

Para um desempenho ideal e para evitar problemas de desempenho, ajuste os parâmetros do kernel que controlam as tabelas de slots TCP.

Executar o `sysctl -a | grep tcp.*.slot_table` comando e respeitar os seguintes parâmetros:

```
# sysctl -a | grep tcp.*.slot_table
sunrpc.tcp_max_slot_table_entries = 128
sunrpc.tcp_slot_table_entries = 128
```

Todos os sistemas Linux devem incluir `sunrpc.tcp_slot_table_entries`, mas apenas alguns incluem `sunrpc.tcp_max_slot_table_entries`. Ambos devem ser definidos para 128.

A falha em definir esses parâmetros pode ter efeitos significativos no desempenho. Em alguns casos, o desempenho é limitado porque o sistema operacional linux não está emitindo e/S suficiente Em outros casos, as latências de e/S aumentam à medida que o sistema operacional linux tenta emitir mais e/S do que pode ser reparado.

Opções de montagem em NFS do Linux

A tabela a seguir lista as opções de montagem NFS do Linux para uma única instância.

Tipo de ficheiro	Opções de montagem
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Arquivos de controle Datafiles Redo logs	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`
ORACLE_HOME	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr`

A tabela a seguir lista as opções de montagem NFS do Linux para RAC.

Tipo de ficheiro	Opções de montagem
ADR Home	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, actimeo=0`
Arquivos de controle arquivos de dados Redo logs	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,actimeo=0`
CRS/votação	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,noac,actimeo=0`
Dedicado `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144`
Compartilhado `ORACLE_HOME`	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,actimeo=0`

A principal diferença entre as opções de montagem de instância única e RAC é a adição `actimeo=0` das opções de montagem. Essa adição tem o efeito de desabilitar o cache do sistema operacional do host, o que permite que todas as instâncias do cluster RAC tenham uma visão consistente do estado dos dados. Embora o uso do `init.ora` parâmetro `filesystemio_options=setall` tenha o mesmo efeito de desabilitar o cache do host, ainda é necessário usar `actimeo=0`o .

O motivo `actimeo=0` necessário para implantações compartilhadas `ORACLE_HOME` é facilitar a consistência de arquivos, como os arquivos de senha Oracle e spfiles. Se cada instância em um cluster RAC tiver um dedicado `ORACLE_HOME`, esse parâmetro não será necessário.

Geralmente, arquivos que não sejam de banco de dados devem ser montados com as mesmas opções usadas para datafiles de instância única, embora aplicativos específicos possam ter requisitos diferentes. Evite as opções de montagem `noac` e `actimeo=0`, se possível, porque essas opções desativam a leitura e o buffer no nível do sistema de arquivos. Isso pode causar problemas graves de desempenho para processos como extração, tradução e carregamento.

ACESSO e GETATTR

Alguns clientes observaram que um nível extremamente alto de outros IOPS, como O ACCESS e GETATTR, pode dominar suas cargas de trabalho. Em casos extremos, operações como leituras e gravações podem ser tão baixas quanto 10% do total. Este é um comportamento normal com qualquer banco de dados que inclua o uso `actimeo=0` e/ou `noac` no Linux porque essas opções fazem com que o sistema operacional Linux recarregue constantemente metadados de arquivos do sistema de armazenamento. Operações como ACCESS e GETATTR são operações de baixo impactos que são atendidas a partir do cache ONTAP em um ambiente de banco de dados. Não devem ser consideradas IOPS originais, como leituras e gravações, que criem verdadeira demanda em sistemas de storage. No entanto, esses outros IOPS criam alguma carga, especialmente em ambientes RAC. Para resolver esta situação, ative o DNFS, que ignora o cache do buffer do sistema operacional e evita essas operações desnecessárias de metadados.

Linux Direct NFS

Uma opção de montagem adicional, chamada `nosharecache`, é necessária quando (a) o DNFS está ativado e (b) um volume de origem é montado mais de uma vez em um único servidor (c) com uma montagem NFS aninhada. Essa configuração é vista principalmente em ambientes compatíveis com aplicações SAP. Por exemplo, um único volume em um sistema NetApp pode ter um diretório localizado em `/vol/oracle/base` e um segundo em `/vol/oracle/home`. Se `/vol/oracle/base` for montado em `/oracle` e `/vol/oracle/home` for montado em `/oracle/home`, o resultado serão montagens NFS aninhadas que se originam na mesma fonte.

O sistema operacional pode detetar o fato de que `/oracle` e `/oracle/home` residir no mesmo volume, que é o mesmo sistema de arquivos de origem. Em seguida, o SO usa o mesmo identificador de dispositivo para acessar os dados. Isso melhora o uso do cache do sistema operacional e outras operações, mas interfere com o DNFS. Se o DNFS tiver de aceder a um ficheiro, como o `spfile`, ligado `/oracle/home`, poderá tentar, erroneamente, utilizar o caminho errado para os dados. O resultado é uma operação de e/S com falha. Nessas configurações, adicione a `nosharecache` opção de montagem a qualquer sistema de arquivos NFS que compartilhe um volume de origem com outro sistema de arquivos NFS nesse host. Isso força o sistema operacional Linux a alocar um identificador de dispositivo independente para esse sistema de arquivos.

Linux Direct NFS e Oracle RAC

O uso do DNFS tem benefícios especiais de desempenho para o Oracle RAC no sistema operacional Linux porque o Linux não tem um método para forçar e/S direto, o que é necessário com RAC para coerência entre os nós. Como solução alternativa, o Linux requer o uso da `actimeo=0` opção de montagem, que faz com que os dados de arquivo expirem imediatamente do cache do sistema operacional. Essa opção, por sua vez, força o cliente NFS Linux a reler constantemente os dados de atributos, o que danifica a latência e aumenta a carga no controlador de armazenamento.

A ativação do DNFS ignora o cliente NFS do host e evita esse dano. Vários clientes relataram melhorias significativas no desempenho em clusters RAC e reduções significativas na carga do ONTAP (especialmente em relação a outros IOPS) ao habilitar o DNFS.

Linux Direct NFS e arquivo oranfstab

Ao usar DNFS no Linux com a opção multipathing, várias sub-redes devem ser usadas. Em outros sistemas operacionais, vários canais DNFS podem ser estabelecidos usando as `LOCAL` opções e `DONTROUTE` para configurar vários canais DNFS em uma única sub-rede. No entanto, isso não funciona corretamente no Linux e problemas de desempenho inesperados podem resultar. Com o Linux, cada NIC usada para o tráfego DNFS deve estar em uma sub-rede diferente.

Programador de e/S.

O kernel Linux permite um controle de baixo nível sobre a maneira como e/S para bloquear dispositivos é agendada. Os padrões em várias distribuições do Linux variam consideravelmente. Testes mostram que o prazo geralmente oferece os melhores resultados, mas ocasionalmente o NOOP foi um pouco melhor. A diferença de desempenho é mínima, mas teste ambas as opções se for necessário extrair o máximo desempenho possível de uma configuração de banco de dados. O CFQ é o padrão em muitas configurações e demonstrou problemas significativos de desempenho com cargas de trabalho de banco de dados.

Consulte a documentação relevante do fornecedor do Linux para obter instruções sobre como configurar o agendador de e/S.

Multipathing

Alguns clientes encontraram falhas durante a interrupção da rede porque o daemon multipath não estava sendo executado em seu sistema. Em versões recentes do Linux, o processo de instalação do sistema operacional e do daemon multipathing podem deixar esses sistemas operacionais vulneráveis a esse problema. Os pacotes são instalados corretamente, mas não são configurados para inicialização automática após uma reinicialização.

Por exemplo, o padrão para o daemon multipath no RHEL5,5 pode aparecer da seguinte forma:

```
[root@host1 iscsi]# chkconfig --list | grep multipath
multipathd      0:off   1:off   2:off   3:off   4:off   5:off   6:off
```

Isso pode ser corrigido com os seguintes comandos:

```
[root@host1 iscsi]# chkconfig multipathd on
[root@host1 iscsi]# chkconfig --list | grep multipath
multipathd      0:off   1:off   2:on    3:on    4:on    5:on    6:off
```

Espelhamento ASM

O espelhamento ASM pode exigir alterações nas configurações de multipath do Linux para permitir que o ASM reconheça um problema e alterne para um grupo de falhas alternativo. A maioria das configurações ASM no ONTAP usa redundância externa, o que significa que a proteção de dados é fornecida pelo array externo e o ASM não espelha dados. Alguns sites usam ASM com redundância normal para fornecer espelhamento bidirecional, normalmente em diferentes sites.

As configurações do Linux mostradas na "Documentação dos utilitários de host do NetApp" incluem parâmetros multipath que resultam em filas indefinidas de e/S Isso significa que uma e/S em um dispositivo LUN sem caminhos ativos aguarda o tempo necessário para que a e/S seja concluída. Isso geralmente é desejável porque os hosts Linux esperam que as alterações de caminho SAN sejam concluídas, que os switches FC sejam reiniciados ou que um sistema de storage conclua um failover.

Esse comportamento ilimitado de enfileiramento causa um problema com o espelhamento ASM porque o ASM deve receber uma falha de e/S para que ele tente novamente e/S em um LUN alternativo.

Defina os seguintes parâmetros no arquivo Linux `multipath.conf` para LUNs ASM usados com espelhamento ASM:

```
polling_interval 5
no_path_retry 24
```

Essas configurações criam um tempo limite de 120 segundos para dispositivos ASM. O tempo limite é calculado como `polling_interval` * `no_path_retry` como segundos. O valor exato pode precisar ser ajustado em algumas circunstâncias, mas um tempo limite de 120 segundos deve ser suficiente para a maioria dos usos. Especificamente, 120 segundos devem permitir que uma tomada de controle ou giveback ocorra sem produzir um erro de e/S que resultaria em que o grupo de falha fosse colocado offline.

Um valor menor `no_path_retry` pode reduzir o tempo necessário para que o ASM alterne para um grupo de falhas alternativo, mas isso também aumenta o risco de um failover indesejado durante atividades de manutenção, como um controle de controle. O risco pode ser atenuado por um monitoramento cuidadoso do estado de espelhamento do ASM. Se ocorrer um failover indesejado, os espelhos podem ser ressinced rapidamente se a ressincronização for executada de forma relativamente rápida. Para obter informações adicionais, consulte a documentação Oracle sobre ASM Fast Mirror Resync para a versão do software Oracle em uso.

Opções de montagem Linux xfs, ext3 e ext4

A NetApp recomenda usando as opções de montagem padrão.

ASMLib/AFD (controlador de filtro ASM)

Tópicos de configuração específicos para o sistema operacional Linux usando AFD e ASMlib

Tamanhos de blocos ASMlib

ASMlib é uma biblioteca de gerenciamento ASM opcional e utilitários associados. Seu valor principal é a capacidade de carimbar um LUN ou um arquivo baseado em NFS como um recurso ASM com uma etiqueta legível por humanos.

Versões recentes do ASMlib detetam um parâmetro LUN chamado Logical Blocks per Physical Block expoent (LBPPBE). Esse valor não foi reportado pelo destino SCSI ONTAP até recentemente. Ele agora retorna um valor que indica que um tamanho de bloco 4KB é preferido. Esta não é uma definição de tamanho de bloco, mas é uma dica para qualquer aplicativo que usa LBPPBE que I/os de um determinado tamanho podem ser manipulados de forma mais eficiente. No entanto, o ASMlib interpreta LBPPBE como um tamanho de bloco e marca persistentemente o cabeçalho ASM quando o dispositivo ASM é criado.

Esse processo pode causar problemas com atualizações e migrações de várias maneiras, tudo com base na incapacidade de misturar dispositivos ASMlib com diferentes tamanhos de bloco no mesmo grupo de discos ASM.

Por exemplo, arrays mais antigos geralmente relataram um valor LBPPBE de 0 ou não relataram esse valor de todo. ASMlib interpreta isso como um tamanho de bloco de 512 bytes. Matrizes mais recentes seriam interpretadas como tendo um tamanho de bloco 4KB. Não é possível misturar dispositivos de 512 bytes e 4KB no mesmo grupo de discos ASM. Isso bloquearia um usuário de aumentar o tamanho do grupo de discos ASM usando LUNs de dois arrays ou utilizando ASM como uma ferramenta de migração. Em outros casos, o RMAN pode não permitir a cópia de arquivos entre um grupo de discos ASM com um tamanho de bloco de 512 bytes e um grupo de discos ASM com um tamanho de bloco de 4KBMB.

A solução preferida é corrigir o ASMlib. O ID do bug Oracle é 13999609, e o patch está presente no oracleasm-support-2,1.8-1 e superior. Este patch permite que um usuário defina o parâmetro `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` como `true` no `/etc/sysconfig/oracleasm` arquivo de configuração. Isso impede que o ASMlib use o parâmetro LBPPBE, o que significa que os LUNs na nova matriz agora são reconhecidos como dispositivos de bloco de 512 bytes.

> A opção não altera o tamanho do bloco em LUNs que foram previamente carimbados pelo ASMlib. Por exemplo, se um grupo de discos ASM com blocos de 512 bytes precisar ser migrado para um novo sistema de armazenamento que relata um bloco 4KB, a opção `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` deve ser definida antes que os novos LUNs sejam carimbados com ASMlib. Se os dispositivos já tiverem sido carimbados por oracleasm, eles devem ser reformatados antes de serem carimbados com um novo tamanho de bloco. Primeiro, desfigure o dispositivo com `oracleasm deletedisk`e, em seguida, limpe os primeiros 1GB do dispositivo com `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Por fim, se o dispositivo tiver sido particionado anteriormente, use o `kpartx` comando para remover partições obsoletas ou simplesmente reiniciar o sistema operacional.

Se o ASMlib não puder ser corrigido, o ASMlib pode ser removido da configuração. Esta alteração é disruptiva e requer a remoção de carimbo de discos ASM e certificar-se de que o `asm_diskstring` parâmetro está definido corretamente. No entanto, essa alteração não requer a migração de dados.

Tamanhos de bloco do Acionamento do filtro ASM (AFD)

O AFD é uma biblioteca de gerenciamento ASM opcional que está se tornando a substituição do ASMlib. Do ponto de vista do armazenamento, ele é muito semelhante ao ASMlib, mas inclui recursos adicionais, como a capacidade de bloquear e/S não-Oracle para reduzir as chances de erros de usuário ou aplicativo que poderiam corromper dados.

Tamanhos de bloco de dispositivos

Como o ASMlib, o AFD também lê o parâmetro LUN blocos lógicos por expoente de bloco físico (LBPPBE) e, por padrão, usa o tamanho do bloco físico, não o tamanho do bloco lógico.

Isso pode criar um problema se AFD for adicionado a uma configuração existente onde os dispositivos ASM já estejam formatados como dispositivos de bloco de 512 bytes. O driver AFD reconheceria o LUN como um dispositivo 4K e a incompatibilidade entre o rótulo ASM e o dispositivo físico impediria o acesso. Da mesma forma, as migrações seriam afetadas porque não é possível misturar dispositivos de 512 bytes e 4KB no mesmo grupo de discos ASM. Isso bloquearia um usuário de aumentar o tamanho do grupo de discos ASM usando LUNs de dois arrays ou utilizando ASM como uma ferramenta de migração. Em outros casos, o RMAN pode não permitir a cópia de arquivos entre um grupo de discos ASM com um tamanho de bloco de 512 bytes e um grupo de discos ASM com um tamanho de bloco de 4KBMB.

A solução é simples - AFD inclui um parâmetro para controlar se usa os tamanhos de blocos lógicos ou físicos. Este é um parâmetro global que afeta todos os dispositivos no sistema. Para forçar o AFD a usar o tamanho do bloco lógico, defina `options oracleafd oracleafd_use_logical_block_size=1` no `/etc/modprobe.d/oracleafd.conf` arquivo.

Tamanhos de transferência multipath

As recentes alterações do kernel do linux impõem restrições de tamanho de e/S enviadas para dispositivos multipath, e o AFD não honra essas restrições. Os I/os são então rejeitados, o que faz com que o caminho LUN fique offline. O resultado é uma incapacidade de instalar o Oracle Grid, configurar ASM ou criar um banco de dados.

A solução é especificar manualmente o comprimento máximo de transferência no arquivo multipath.conf para LUNs ONTAP:

```
devices {
            device {
                vendor "NETAPP"
                product "LUN.*"
                max_sectors_kb 4096
        }
    }
```

Mesmo que não existam problemas atualmente, este parâmetro deve ser definido se AFD for usado para garantir que uma futura atualização do linux não cause problemas inesperadamente.

Microsoft Windows

Tópicos de configuração para banco de dados Oracle no Microsoft Windows com ONTAP.

NFS

A Oracle suporta o uso do Microsoft Windows com o cliente NFS direto. Esse recurso oferece um caminho para os benefícios de gerenciamento do NFS, incluindo a capacidade de exibir arquivos entre ambientes, redimensionar volumes dinamicamente e utilizar um protocolo IP menos caro. Consulte a documentação oficial da Oracle para obter informações sobre como instalar e configurar um banco de dados no Microsoft Windows usando DNFS. Não existem boas práticas especiais.

SAN

Para uma ótima eficiência de compressão, certifique-se de que o sistema de ficheiros NTFS utilize uma unidade de alocação de 8K GB ou maior. O uso de uma unidade de alocação 4K, que geralmente é o padrão, afeta negativamente a eficiência da compressão.

Solaris

Tópicos de configuração específicos do Solaris os.

Opções de montagem do Solaris NFS

A tabela a seguir lista as opções de montagem do Solaris NFS para uma única instância.

Tipo de ficheiro	Opções de montagem
ADR Home	`rw,bg,hard,[vers=3,vers=4.1], roto=tcp, timeo=600,rsize=262144,wsize=262144`
Registros do Redo ControlFiles Datafiles	`rw,bg,hard,[vers=3,vers=4.1],proto=tcp, timeo=600,rsize=262144,wsize=262144, nointr,llock,suid`

Tipo de ficheiro	Opções de montagem
ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid

Provou-se que o uso do `llock` melhora significativamente a performance nos ambientes dos clientes, eliminando a latência associada à aquisição e liberação de bloqueios no sistema de storage. Use essa opção com cuidado em ambientes nos quais vários servidores são configurados para montar os mesmos sistemas de arquivos e o Oracle está configurado para montar esses bancos de dados. Embora esta seja uma configuração altamente incomum, ela é usada por um pequeno número de clientes. Se uma instância for iniciada acidentalmente uma segunda vez, a corrupção de dados pode ocorrer porque a Oracle não consegue detetar os arquivos de bloqueio no servidor estrangeiro. Os bloqueios NFS não oferecem proteção de outra forma; como no NFS versão 3, eles são apenas consultivos.

Como `llock` os parâmetros e `forcedirectio` são mutuamente exclusivos, é importante que `filesystemio_options=setall` esteja presente no `init.ora` arquivo para que `directio` seja usado. Sem esse parâmetro, o cache do buffer do sistema operacional do host é usado e o desempenho pode ser afetado negativamente.

A tabela a seguir lista as opções de montagem do Solaris NFS RAC.

Tipo de ficheiro	Opções de montagem
ADR Home	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,noac
Arquivos de controle arquivos de dados Redo logs	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio
CRS/votação	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,forcedirectio
Dedicado ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,suid
Compartilhado ORACLE_HOME	rw,bg,hard,[vers=3,vers=4.1],proto=tcp,timeo=600,rsize=262144,wsize=262144,nointr,noac,suid

A principal diferença entre as opções de montagem de uma única instância e RAC é a adição `noac` de e `forcedirectio` às opções de montagem. Essa adição tem o efeito de desabilitar o cache do sistema operacional do host, o que permite que todas as instâncias do cluster RAC tenham uma visão consistente do estado dos dados. Embora o uso do `init.ora` parâmetro `filesystemio_options=setall` tenha o mesmo efeito de desabilitar o cache do host, ainda é necessário usar `noac` e `forcedirectio`.

O motivo `actimeo=0` necessário para implantações compartilhadas `ORACLE_HOME` é facilitar a consistência de arquivos, como arquivos de senha Oracle e spfiles. Se cada instância em um cluster RAC tiver um dedicado `ORACLE_HOME`, esse parâmetro não será necessário.

Opções de montagem do Solaris UFS

A NetApp recomenda fortemente o uso da opção de montagem de log para que a integridade dos dados seja preservada no caso de uma falha de host do Solaris ou a interrupção da conetividade FC. A opção de montagem de log também preserva a usabilidade dos backups Snapshot.

Solaris ZFS

O Solaris ZFS deve ser instalado e configurado cuidadosamente para oferecer o melhor desempenho.

mvector

O Solaris 11 incluiu uma mudança na forma como processa grandes operações de e/S, o que pode resultar em graves problemas de desempenho em matrizes de armazenamento SAN. O problema está documentado no relatório de bug de rastreamento do NetApp 630173, "regressão de desempenho do Solaris 11 ZFS".

Este não é um bug do ONTAP. É um defeito do Solaris que é rastreado sob os defeitos Solaris 7199305 e 7082975.

Você pode consultar o suporte Oracle para saber se sua versão do Solaris 11 é afetada ou testar a solução alternativa alterando `zfs_mvector_max_size` para um valor menor.

Você pode fazer isso executando o seguinte comando como root:

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t131072" |mdb -kw
```

Se surgir algum problema inesperado dessa alteração, ela pode ser facilmente revertida executando o seguinte comando como root:

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t1048576" |mdb -kw
```

Kernel

O desempenho confiável do ZFS requer um kernel Solaris corrigido contra problemas de alinhamento de LUN. A correção foi introduzida com o patch 147440-19 no Solaris 10 e com o SRU 10,5 para Solaris 11. Utilize apenas o Solaris 10 e posterior com o ZFS.

Configuração LUN

Para configurar um LUN, execute as seguintes etapas:

1. Crie um LUN do tipo `solaris`.
2. Instale o Kit de Utilitário do host (HUK) apropriado especificado pelo "Ferramenta de Matriz de interoperabilidade NetApp (IMT)".
3. Siga as instruções no HUK exatamente como descrito. Os passos básicos estão descritos abaixo, mas consulte o "documentação mais recente" para obter o procedimento adequado.
    a. Execute o `host_config` utilitário para atualizar o `sd.conf/sdd.conf` arquivo. Isso permite que as unidades SCSI descubram corretamente LUNs ONTAP.
    b. Siga as instruções dadas pelo `host_config` utilitário para ativar a entrada/saída multipath (MPIO).

c. Reinicie. Esta etapa é necessária para que quaisquer alterações sejam reconhecidas em todo o sistema.

4. Particione os LUNs e verifique se eles estão alinhados corretamente. Consulte o "Apêndice B: Verificação do alinhamento do WAFL" para obter instruções sobre como testar e confirmar diretamente o alinhamento.

zpools

Um zpool só deve ser criado após as etapas no "Configuração LUN" serem executadas. Se o procedimento não for feito corretamente, pode resultar em degradação grave do desempenho devido ao alinhamento de e/S. O desempenho ideal no ONTAP requer que a e/S seja alinhada a um limite de 4K mm numa unidade. Os sistemas de arquivos criados em um zpool usam um tamanho de bloco efetivo que é controlado por meio de um parâmetro `ashift` chamado , que pode ser visualizado executando o comando `zdb -C`.

O valor `ashift` padrão é 9, o que significa 2 9, ou 512 bytes. Para um desempenho ideal, o `ashift` valor deve ser 12 (2 12 4K). Esse valor é definido no momento em que o zpool é criado e não pode ser alterado, o que significa que os dados em zpools com `ashift` outros que não 12 devem ser migrados copiando dados para um zpool recém-criado.

Depois de criar um zpool, verifique o valor de `ashift` antes de continuar. Se o valor não for 12, os LUNs não foram detetados corretamente. Destrua o zpool, verifique se todas as etapas mostradas na documentação relevante dos Utilitários do host foram executadas corretamente e recrie o zpool.

Zpools e Solaris LDOMs

Os Solaris LDOMs criam um requisito adicional para garantir que o alinhamento de e/S esteja correto. Embora um LUN possa ser encontrado corretamente como um dispositivo 4K, um dispositivo vdsk virtual em um LDOM não herda a configuração do domínio de e/S. O vdsk baseado nesse LUN retorna para um bloco de 512 bytes.

É necessário um ficheiro de configuração adicional. Primeiro, os LDOM individuais devem ser corrigidos para o bug Oracle 15824910 para habilitar as opções de configuração adicionais. Este patch foi portado para todas as versões usadas atualmente do Solaris. Uma vez que o LDOM é corrigido, ele está pronto para a configuração dos novos LUNs corretamente alinhados da seguinte forma:

1. Identifique o LUN ou LUNs a serem usados no novo zpool. Neste exemplo, é o dispositivo c2d1.

```
[root@LDOM1 ~]# echo | format
Searching for disks...done
AVAILABLE DISK SELECTIONS:
  0. c2d0 <Unknown-Unknown-0001-100.00GB>
     /virtual-devices@100/channel-devices@200/disk@0
  1. c2d1 <SUN-ZFS Storage 7330-1.0 cyl 1623 alt 2 hd 254 sec 254>
     /virtual-devices@100/channel-devices@200/disk@1
```

2. Recupere a instância vdc dos dispositivos a serem usados para um pool ZFS:

```
[root@LDOM1 ~]#  cat /etc/path_to_inst
#
# Caution! This file contains critical kernel state
#
"/fcoe" 0 "fcoe"
"/iscsi" 0 "iscsi"
"/pseudo" 0 "pseudo"
"/scsi_vhci" 0 "scsi_vhci"
"/options" 0 "options"
"/virtual-devices@100" 0 "vnex"
"/virtual-devices@100/channel-devices@200" 0 "cnex"
"/virtual-devices@100/channel-devices@200/disk@0" 0 "vdc"
"/virtual-devices@100/channel-devices@200/pciv-communication@0" 0 "vpci"
"/virtual-devices@100/channel-devices@200/network@0" 0 "vnet"
"/virtual-devices@100/channel-devices@200/network@1" 1 "vnet"
"/virtual-devices@100/channel-devices@200/network@2" 2 "vnet"
"/virtual-devices@100/channel-devices@200/network@3" 3 "vnet"
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc" << We want
this one
```

3. `/platform/sun4v/kernel/drv/vdc.conf`Editar :

```
block-size-list="1:4096";
```

Isso significa que a instância 1 do dispositivo recebe um tamanho de bloco de 4096MB.

Como exemplo adicional, suponha que as instâncias 1 a 6 do vdsk precisem ser configuradas para um tamanho de bloco 4K e `/etc/path_to_inst` lerem da seguinte forma:

```
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc"
"/virtual-devices@100/channel-devices@200/disk@2" 2 "vdc"
"/virtual-devices@100/channel-devices@200/disk@3" 3 "vdc"
"/virtual-devices@100/channel-devices@200/disk@4" 4 "vdc"
"/virtual-devices@100/channel-devices@200/disk@5" 5 "vdc"
"/virtual-devices@100/channel-devices@200/disk@6" 6 "vdc"
```

4. O arquivo final `vdc.conf` deve conter o seguinte:

```
block-size-list="1:8192","2:8192","3:8192","4:8192","5:8192","6:8192";
```

Cuidado
O LDOM deve ser reinicializado depois que o vdc.conf é configurado e o vdsk é criado. Este passo não pode ser evitado. A alteração do tamanho do bloco só entra em vigor após uma reinicialização. Prossiga com a configuração do zpool e certifique-se de que o ashift está corretamente configurado para 12, conforme descrito anteriormente.

Registo intenção ZFS (ZIL)

Geralmente, não há razão para localizar o ZFS Intent Log (ZIL) em um dispositivo diferente. O log pode compartilhar espaço com a piscina principal. O uso principal de um ZIL separado é quando se usa unidades físicas que não têm os recursos de armazenamento em cache de gravação em arrays de armazenamento modernos.

logbias

Defina `logbias` o parâmetro em sistemas de arquivos ZFS que hospedam dados Oracle.

```
zfs set logbias=throughput <filesystem>
```

O uso desse parâmetro reduz os níveis gerais de gravação. Sob os padrões, os dados escritos são comprometidos primeiro com o ZIL e depois para o pool de armazenamento principal. Essa abordagem é apropriada para uma configuração usando uma configuração de unidade simples, que inclui um dispositivo ZIL baseado em SSD e Mídia giratória para o pool de armazenamento principal. Isso ocorre porque permite que um commit ocorra em uma única transação de e/S na Mídia de menor latência disponível.

Ao usar um storage array moderno que inclua sua própria funcionalidade de armazenamento em cache, essa abordagem geralmente não é necessária. Em circunstâncias raras, pode ser desejável submeter uma gravação com uma única transação ao log, como uma carga de trabalho que consiste em gravações aleatórias altamente concentradas e sensíveis à latência. Há consequências na forma de amplificação de gravação porque os dados registrados são gravados no pool de armazenamento principal, resultando em uma duplicação da atividade de gravação.

E/S direta

Muitas aplicações, incluindo produtos Oracle, podem ignorar o cache de buffer do host habilitando a e/S direta Esta estratégia não funciona como esperado com sistemas de arquivos ZFS. Embora o cache do buffer do host seja ignorado, o próprio ZFS continua a armazenar dados em cache. Essa ação pode resultar em resultados enganosos ao usar ferramentas como fio ou sio para realizar testes de desempenho, pois é difícil prever se a e/S está chegando ao sistema de armazenamento ou se está sendo armazenada em cache localmente no sistema operacional. Essa ação também torna muito difícil usar esses testes sintéticos para comparar o desempenho do ZFS com outros sistemas de arquivos. Na prática, há pouca ou nenhuma diferença no desempenho do sistema de arquivos em workloads reais do usuário.

Vários zpools

Backups, restaurações, clones e arquivamento baseados em snapshot de dados baseados em ZFS devem ser executados no nível do zpool e, geralmente, exigem vários zpools. Um zpool é análogo a um grupo de discos LVM e deve ser configurado usando as mesmas regras. Por exemplo, um banco de dados provavelmente é melhor definido com os datafiles que residem em `zpool1` e os logs de arquivo, arquivos de controle e logs de refazer que residem em `zpool2`. Essa abordagem permite um hot backup padrão no qual o banco de dados é colocado no modo hot backup, seguido por um snapshot de `zpool1`. O banco de dados é

então removido do modo hot backup, o arquivo de log é forçado e um snapshot de `zpool2` é criado. Uma operação de restauração requer a desmontagem dos sistemas de arquivos zfs e a remoção do zpool em sua totalidade, seguindo-se uma operação de restauração do SnapRestore. O zpool pode então ser colocado on-line novamente e o banco de dados recuperado.

sistema de arquivos_options

O parâmetro Oracle `filesystemio_options` funciona de forma diferente com o ZFS. Se `setall` ou `directio` for usado, as operações de gravação são síncronas e ignoram o cache do buffer do sistema operacional, mas as leituras são armazenadas em buffer pelo ZFS. Essa ação causa dificuldades na análise de desempenho, pois às vezes, a e/S é intercetada e atendida pelo cache ZFS, tornando a latência do armazenamento e a e/S total menos do que parece ser.

Configuração de host com sistemas ASA r2

AIX

Tópicos de configuração para banco de dados Oracle no IBM AIX com ASA r2 ONTAP.

> O AIX é compatível com o NetApp ASA r2 para hospedagem de bancos de dados Oracle, desde que:
>
> - Você configurou o Oracle corretamente para E/S simultânea.
> - Você utiliza protocolos SAN compatíveis (FC/iSCSI/NVMe).
> - Você executa o ONTAP 9.16.x ou posterior no ASA r2.

E/S simultânea

Para obter o desempenho ideal no IBM AIX com ASA r2, é necessário utilizar E/S simultânea. Sem E/S simultânea, é provável que haja limitações de desempenho, pois o AIX realiza E/S serializada e atômica, o que acarreta uma sobrecarga significativa.

Originalmente, a NetApp recomendava o uso do `cio` A opção de montagem forçava a E/S simultânea no sistema de arquivos, mas esse processo tinha desvantagens e não é mais necessário. Desde a introdução do AIX 5.2 e do Oracle 10gR1, o Oracle no AIX pode abrir arquivos individuais para E/S simultânea, em vez de forçar E/S simultânea em todo o sistema de arquivos.

O melhor método para ativar e/S concorrente é definir o `init.ora` parâmetro `filesystemio_options` como `setall`. Isso permite que a Oracle abra arquivos específicos para uso com e/S concorrente

Usar o cio como opção de montagem força o uso de E/S simultânea, o que pode ter consequências negativas. Por exemplo, forçar E/S simultânea desativa a leitura antecipada nos sistemas de arquivos, o que pode prejudicar o desempenho de E/S que ocorrem fora do software de banco de dados Oracle, como copiar arquivos e realizar backups em fita. Além disso, produtos como o Oracle GoldenGate e o SAP BR*Tools não são compatíveis com o uso da opção de montagem cio em determinadas versões do Oracle.

A NetApp recomenda o seguinte:

- Não use a `cio` opção de montagem no nível do sistema de arquivos. Em vez disso, ative a e/S concorrente através do uso `filesystemio_options=setall` do .
- Use apenas o `cio` opção de montagem se não for possível definir `filesystemio_options=setall`.

Como o ASA r2 não suporta NAS, todas as implementações do Oracle no AIX devem usar protocolos de bloco.

AIX jfs/jfs2 Opções de montagem

A tabela a seguir lista as opções de montagem do AIX jfs/jfs2.

Tipo de ficheiro	Opções de montagem
ADR Home	Predefinições
Arquivos de controle	Predefinições
Arquivos de dados	Predefinições
Logs de refazer	Predefinições
ORACLE_HOME	Predefinições

Antes de usar o AIX `hdisk` dispositivos em qualquer ambiente, incluindo bancos de dados, verifique o parâmetro `queue_depth`. Este parâmetro não se refere à profundidade da fila HBA; em vez disso, relaciona-se à profundidade da fila SCSI do dispositivo individual. `hdisk device`. Dependendo de como os LUNs do ASA r2 estão configurados, o valor para `queue_depth` Pode ser muito baixo para um bom desempenho. Os testes demonstraram que o valor ideal é 64.

HP-UX

Tópicos de configuração para banco de dados Oracle em HP-UX com ASA r2 ONTAP.

O HP-UX é compatível com o NetApp ASA r2 para hospedagem de bancos de dados Oracle, desde que:

- A versão do ONTAP é 9.16.x ou posterior.
- Utilize protocolos SAN (FC/iSCSI/NVMe). O NAS não é suportado no ASA r2.
- Aplique as melhores práticas de montagem e ajuste de E/S específicas para HP-UX.

Opções de montagem HP-UX VxFS

Use as seguintes opções de montagem para sistemas de arquivos que hospedam binários Oracle:

```
delaylog,nodatainlog
```

Use as seguintes opções de montagem para sistemas de arquivos que contêm datafiles, logs de refazer, logs

de arquivamento e arquivos de controle nos quais a versão do HP-UX não suporta e/S simultânea:

```
nodatainlog,mincache=direct,convosync=direct
```

Quando a e/S simultânea for suportada (VxFS 5.0.1 e posterior, ou com o ServiceGuard Storage Management Suite), utilize estas opções de montagem para sistemas de ficheiros que contenham ficheiros de dados, registos refeitos, registos de arquivo e ficheiros de controlo:

```
delaylog,cio
```

> O parâmetro `db_file_multiblock_read_count` é especialmente crítico em ambientes VxFS. A Oracle recomenda que esse parâmetro permaneça desconfigurado no Oracle 10g R1i e posterior, a menos que especificamente direcionado de outra forma. O padrão com um tamanho de bloco Oracle 8KB é 128. Se o valor deste parâmetro for forçado a 16 ou menos, remova a `convosync=direct` opção de montagem porque pode danificar o desempenho sequencial de e/S. Esta etapa prejudica outros aspetos do desempenho e só deve ser tomada se o valor de `db_file_multiblock_read_count` tiver de ser alterado do valor padrão.

Linux

Tópicos de configuração específicos para o sistema operacional Linux com ASA r2 ONTAP.

> O Linux (Oracle Linux, RHEL, SUSE) é compatível com o ASA r2 para bancos de dados Oracle. Utilize protocolos SAN, configure o multipathing corretamente e aplique as melhores práticas da Oracle para otimização de ASM e E/S.

Programador de e/S.

O kernel Linux permite um controle de baixo nível sobre a maneira como e/S para bloquear dispositivos é agendada. Os padrões em várias distribuições do Linux variam consideravelmente. Testes mostram que o prazo geralmente oferece os melhores resultados, mas ocasionalmente o NOOP foi um pouco melhor. A diferença de desempenho é mínima, mas teste ambas as opções se for necessário extrair o máximo desempenho possível de uma configuração de banco de dados. O CFQ é o padrão em muitas configurações e demonstrou problemas significativos de desempenho com cargas de trabalho de banco de dados.

Consulte a documentação relevante do fornecedor do Linux para obter instruções sobre como configurar o agendador de e/S.

Multipathing

Alguns clientes encontraram falhas durante a interrupção da rede porque o daemon multipath não estava sendo executado em seu sistema. Em versões recentes do Linux, o processo de instalação do sistema operacional e do daemon multipathing podem deixar esses sistemas operacionais vulneráveis a esse problema. Os pacotes são instalados corretamente, mas não são configurados para inicialização automática após uma reinicialização.

Por exemplo, a configuração padrão do daemon multipath no RHEL 9.7 pode ser a seguinte:

```
[root@host1 ~]# systemctl list-unit-files --type=service | grep multipathd
multipathd.service                        disabled
```

Isso pode ser corrigido com os seguintes comandos:

```
[root@host1 ~]# systemctl enable multipathd.service
[root@host1 ~]# systemctl list-unit-files --type=service | grep multipathd
multipathd.service                        enabled
```

Profundidade da fila

Defina a profundidade de fila adequada para os dispositivos SAN a fim de evitar gargalos de E/S. A profundidade de fila padrão no Linux geralmente é definida como 128, o que pode causar problemas de desempenho com bancos de dados Oracle. Configurar uma profundidade de fila muito alta pode causar enfileiramento excessivo de operações de E/S, levando ao aumento da latência e à redução da taxa de transferência. Definir um valor muito baixo pode limitar o número de solicitações de E/S pendentes, reduzindo o desempenho geral. Uma profundidade de fila de 64 costuma ser um bom ponto de partida para cargas de trabalho de banco de dados Oracle no ASA r2, mas pode precisar ser ajustada com base nas características específicas da carga de trabalho e nos testes de desempenho.

Espelhamento ASM

O espelhamento ASM pode exigir alterações nas configurações de multipath do Linux para permitir que o ASM reconheça um problema e alterne para um grupo de falhas alternativo. A maioria das configurações ASM no ONTAP usa redundância externa, o que significa que a proteção de dados é fornecida pelo array externo e o ASM não espelha dados. Alguns sites usam ASM com redundância normal para fornecer espelhamento bidirecional, normalmente em diferentes sites.

Para sistemas ASA r2 que suportam multipathing ativo-ativo, essas configurações de multipathing devem ser ajustadas. Como todos os caminhos estão ativos e com balanceamento de carga, o enfileiramento indefinido não é necessário. Em vez disso, os parâmetros de múltiplos caminhos devem priorizar o desempenho e a rápida recuperação em caso de falha. Esse comportamento é importante para o espelhamento ASM, pois o ASM precisa receber uma falha de E/S para tentar novamente a E/S em um LUN alternativo. Se as operações de E/S forem enfileiradas indefinidamente, o ASM não poderá acionar um failover.

Defina os seguintes parâmetros no arquivo Linux `multipath.conf` para LUNs ASM usados com espelhamento ASM:

```
polling_interval 5
no_path_retry 24
failback immediate
path_grouping_policy multibus
path_selector "service-time 0"
```

Essas configurações criam um tempo limite de 120 segundos para dispositivos ASM. O tempo limite é calculado como `polling_interval` * `no_path_retry` como segundos. O valor exato pode precisar ser ajustado em algumas circunstâncias, mas um tempo limite de 120 segundos deve ser suficiente para a maioria dos usos. Especificamente, 120 segundos devem permitir que uma tomada de controle ou giveback

ocorra sem produzir um erro de e/S que resultaria em que o grupo de falha fosse colocado offline.

Um valor menor `no_path_retry` pode reduzir o tempo necessário para que o ASM alterne para um grupo de falhas alternativo, mas isso também aumenta o risco de um failover indesejado durante atividades de manutenção, como um controle de controle. O risco pode ser atenuado por um monitoramento cuidadoso do estado de espelhamento do ASM. Se ocorrer um failover indesejado, os espelhos podem ser ressinced rapidamente se a ressincronização for executada de forma relativamente rápida. Para obter informações adicionais, consulte a documentação Oracle sobre ASM Fast Mirror Resync para a versão do software Oracle em uso.

Opções de montagem Linux xfs, ext3 e ext4

* A NetApp recomenda* o uso das opções de montagem padrão. Ao criar sistemas de arquivos em LUNs, assegure-se de que o alinhamento esteja correto.

ASMLib/AFD (controlador de filtro ASM)

Tópicos de configuração específicos para o sistema operacional Linux usando AFD e ASMlib com ASA r2 ONTAP.

Tamanhos de blocos ASMlib

ASMlib é uma biblioteca opcional de gerenciamento do ASM e utilitários associados. Sua principal vantagem reside na capacidade de marcar um LUN como um recurso ASM com um rótulo legível por humanos.

Versões recentes do ASMlib detetam um parâmetro LUN chamado Logical Blocks per Physical Block expoent (LBPPBE). Esse valor não foi reportado pelo destino SCSI ONTAP até recentemente. Ele agora retorna um valor que indica que um tamanho de bloco 4KB é preferido. Esta não é uma definição de tamanho de bloco, mas é uma dica para qualquer aplicativo que usa LBPPBE que I/os de um determinado tamanho podem ser manipulados de forma mais eficiente. No entanto, o ASMlib interpreta LBPPBE como um tamanho de bloco e marca persistentemente o cabeçalho ASM quando o dispositivo ASM é criado.

Esse processo pode causar problemas com atualizações e migrações de várias maneiras, tudo com base na incapacidade de misturar dispositivos ASMlib com diferentes tamanhos de bloco no mesmo grupo de discos ASM.

Por exemplo, arrays mais antigos geralmente relataram um valor LBPPBE de 0 ou não relataram esse valor de todo. ASMlib interpreta isso como um tamanho de bloco de 512 bytes. Matrizes mais recentes seriam interpretadas como tendo um tamanho de bloco 4KB. Não é possível misturar dispositivos de 512 bytes e 4KB no mesmo grupo de discos ASM. Isso bloquearia um usuário de aumentar o tamanho do grupo de discos ASM usando LUNs de dois arrays ou utilizando ASM como uma ferramenta de migração. Em outros casos, o RMAN pode não permitir a cópia de arquivos entre um grupo de discos ASM com um tamanho de bloco de 512 bytes e um grupo de discos ASM com um tamanho de bloco de 4KBMB.

A solução preferida é corrigir o ASMlib. O ID do bug Oracle é 13999609, e o patch está presente no oracleasm-support-2,1.8-1 e superior. Este patch permite que um usuário defina o parâmetro `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` como `true` no `/etc/sysconfig/oracleasm` arquivo de configuração. Isso impede que o ASMlib use o parâmetro LBPPBE, o que significa que os LUNs na nova matriz agora são reconhecidos como dispositivos de bloco de 512 bytes.

A opção não altera o tamanho do bloco em LUNs que foram previamente carimbados pelo ASMlib. Por exemplo, se um grupo de discos ASM com blocos de 512 bytes precisar ser migrado para um novo sistema de armazenamento que relata um bloco 4KB, a opção `ORACLEASM_USE_LOGICAL_BLOCK_SIZE` deve ser definida antes que os novos LUNs sejam carimbados com ASMlib. Se os dispositivos já tiverem sido carimbados por oracleasm, eles devem ser reformatados antes de serem carimbados com um novo tamanho de bloco. Primeiro, desfigure o dispositivo com `oracleasm deletedisk`e, em seguida, limpe os primeiros 1GB do dispositivo com `dd if=/dev/zero of=/dev/mapper/device bs=1048576 count=1024`. Por fim, se o dispositivo tiver sido particionado anteriormente, use o `kpartx` comando para remover partições obsoletas ou simplesmente reiniciar o sistema operacional.

Se o ASMlib não puder ser corrigido, o ASMlib pode ser removido da configuração. Esta alteração é disruptiva e requer a remoção de carimbo de discos ASM e certificar-se de que o `asm_diskstring` parâmetro está definido corretamente. No entanto, essa alteração não requer a migração de dados.

Tamanhos de bloco do Acionamento do filtro ASM (AFD)

O AFD é uma biblioteca de gerenciamento ASM opcional que está se tornando a substituição do ASMlib. Do ponto de vista do armazenamento, ele é muito semelhante ao ASMlib, mas inclui recursos adicionais, como a capacidade de bloquear e/S não-Oracle para reduzir as chances de erros de usuário ou aplicativo que poderiam corromper dados.

Tamanhos de bloco de dispositivos

Como o ASMlib, o AFD também lê o parâmetro LUN blocos lógicos por expoente de bloco físico (LBPPBE) e, por padrão, usa o tamanho do bloco físico, não o tamanho do bloco lógico.

Isso pode criar um problema se AFD for adicionado a uma configuração existente onde os dispositivos ASM já estejam formatados como dispositivos de bloco de 512 bytes. O driver AFD reconheceria o LUN como um dispositivo 4K e a incompatibilidade entre o rótulo ASM e o dispositivo físico impediria o acesso. Da mesma forma, as migrações seriam afetadas porque não é possível misturar dispositivos de 512 bytes e 4KB no mesmo grupo de discos ASM. Isso bloquearia um usuário de aumentar o tamanho do grupo de discos ASM usando LUNs de dois arrays ou utilizando ASM como uma ferramenta de migração. Em outros casos, o RMAN pode não permitir a cópia de arquivos entre um grupo de discos ASM com um tamanho de bloco de 512 bytes e um grupo de discos ASM com um tamanho de bloco de 4KBMB.

A solução é simples - AFD inclui um parâmetro para controlar se usa os tamanhos de blocos lógicos ou físicos. Este é um parâmetro global que afeta todos os dispositivos no sistema. Para forçar o AFD a usar o tamanho do bloco lógico, defina `options oracleafd oracleafd_use_logical_block_size=1` no `/etc/modprobe.d/oracleafd.conf` arquivo.

Tamanhos de transferência multipath

As recentes alterações do kernel do linux impõem restrições de tamanho de e/S enviadas para dispositivos multipath, e o AFD não honra essas restrições. Os I/os são então rejeitados, o que faz com que o caminho LUN fique offline. O resultado é uma incapacidade de instalar o Oracle Grid, configurar ASM ou criar um banco de dados.

A solução é especificar manualmente o comprimento máximo de transferência no arquivo multipath.conf para LUNs ONTAP:

```
devices {
            device {
                vendor "NETAPP"
                product "LUN.*"
                max_sectors_kb 4096
        }
    }
```

Mesmo que não existam problemas atualmente, este parâmetro deve ser definido se AFD for usado para garantir que uma futura atualização do linux não cause problemas inesperadamente.

Microsoft Windows

Tópicos de configuração para banco de dados Oracle no Microsoft Windows com ASA r2 ONTAP.

SAN

Para uma ótima eficiência de compressão, certifique-se de que o sistema de ficheiros NTFS utilize uma unidade de alocação de 8K GB ou maior. O uso de uma unidade de alocação 4K, que geralmente é o padrão, afeta negativamente a eficiência da compressão.

Solaris

Tópicos de configuração específicos para o sistema operacional Solaris com ASA r2 ONTAP.

Opções de montagem do Solaris UFS

A NetApp recomenda fortemente o uso da opção de montagem de log para que a integridade dos dados seja preservada no caso de uma falha de host do Solaris ou a interrupção da conetividade FC. A opção de montagem de log também preserva a usabilidade dos backups Snapshot.

Solaris ZFS

O Solaris ZFS deve ser instalado e configurado cuidadosamente para oferecer o melhor desempenho.

mvector

O Solaris 11 incluiu uma mudança na forma como processa grandes operações de e/S, o que pode resultar em graves problemas de desempenho em matrizes de armazenamento SAN. O problema está documentado no relatório de bug de rastreamento do NetApp 630173, "regressão de desempenho do Solaris 11 ZFS".

Este não é um bug do ONTAP. É um defeito do Solaris que é rastreado sob os defeitos Solaris 7199305 e 7082975.

Você pode consultar o suporte Oracle para saber se sua versão do Solaris 11 é afetada ou testar a solução alternativa alterando `zfs_mvector_max_size` para um valor menor.

Você pode fazer isso executando o seguinte comando como root:

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t131072" |mdb -kw
```

Se surgir algum problema inesperado dessa alteração, ela pode ser facilmente revertida executando o seguinte comando como root:

```
[root@host1 ~]# echo "zfs_mvector_max_size/W 0t1048576" |mdb -kw
```

Kernel

O desempenho confiável do ZFS requer um kernel Solaris corrigido contra problemas de alinhamento de LUN. A correção foi introduzida com o patch 147440-19 no Solaris 10 e com o SRU 10,5 para Solaris 11. Utilize apenas o Solaris 10 e posterior com o ZFS.

Configuração LUN

Para configurar um LUN, execute as seguintes etapas:

1. Crie um LUN do tipo `solaris`.
2. Instale o Kit de Utilitário do host (HUK) apropriado especificado pelo "Ferramenta de Matriz de interoperabilidade NetApp (IMT)".
3. Siga as instruções no HUK exatamente como descrito. Os passos básicos estão descritos abaixo, mas consulte o "documentação mais recente" para obter o procedimento adequado.
    a. Execute o `host_config` utilitário para atualizar o `sd.conf/sdd.conf` arquivo. Isso permite que as unidades SCSI descubram corretamente LUNs ONTAP.
    b. Siga as instruções dadas pelo `host_config` utilitário para ativar a entrada/saída multipath (MPIO).
    c. Reinicie. Esta etapa é necessária para que quaisquer alterações sejam reconhecidas em todo o sistema.
4. Particione os LUNs e verifique se eles estão alinhados corretamente. Consulte o "Apêndice B: Verificação do alinhamento do WAFL" para obter instruções sobre como testar e confirmar diretamente o alinhamento.

zpools

Um zpool só deve ser criado após as etapas no "Configuração LUN" serem executadas. Se o procedimento não for feito corretamente, pode resultar em degradação grave do desempenho devido ao alinhamento de e/S. O desempenho ideal no ONTAP requer que a e/S seja alinhada a um limite de 4K mm numa unidade. Os sistemas de arquivos criados em um zpool usam um tamanho de bloco efetivo que é controlado por meio de um parâmetro `ashift` chamado , que pode ser visualizado executando o comando `zdb -C`.

O valor `ashift` padrão é 9, o que significa 2 9, ou 512 bytes. Para um desempenho ideal, o `ashift` valor deve ser 12 (2 12 4K). Esse valor é definido no momento em que o zpool é criado e não pode ser alterado, o que significa que os dados em zpools com `ashift` outros que não 12 devem ser migrados copiando dados para um zpool recém-criado.

Depois de criar um zpool, verifique o valor de `ashift` antes de continuar. Se o valor não for 12, os LUNs não foram detetados corretamente. Destrua o zpool, verifique se todas as etapas mostradas na documentação relevante dos Utilitários do host foram executadas corretamente e recrie o zpool.

Zpools e Solaris LDOMs

Os Solaris LDOMs criam um requisito adicional para garantir que o alinhamento de e/S esteja correto. Embora um LUN possa ser encontrado corretamente como um dispositivo 4K, um dispositivo vdsk virtual em um LDOM não herda a configuração do domínio de e/S. O vdsk baseado nesse LUN retorna para um bloco de 512 bytes.

É necessário um ficheiro de configuração adicional. Primeiro, os LDOM individuais devem ser corrigidos para o bug Oracle 15824910 para habilitar as opções de configuração adicionais. Este patch foi portado para todas as versões usadas atualmente do Solaris. Uma vez que o LDOM é corrigido, ele está pronto para a configuração dos novos LUNs corretamente alinhados da seguinte forma:

1. Identifique o LUN ou LUNs a serem usados no novo zpool. Neste exemplo, é o dispositivo c2d1.

```
[root@LDOM1 ~]# echo | format
Searching for disks...done
AVAILABLE DISK SELECTIONS:
  0. c2d0 <Unknown-Unknown-0001-100.00GB>
     /virtual-devices@100/channel-devices@200/disk@0
  1. c2d1 <SUN-ZFS Storage 7330-1.0 cyl 1623 alt 2 hd 254 sec 254>
     /virtual-devices@100/channel-devices@200/disk@1
```

2. Recupere a instância vdc dos dispositivos a serem usados para um pool ZFS:

```
[root@LDOM1 ~]#  cat /etc/path_to_inst
#
# Caution! This file contains critical kernel state
#
"/fcoe" 0 "fcoe"
"/iscsi" 0 "iscsi"
"/pseudo" 0 "pseudo"
"/scsi_vhci" 0 "scsi_vhci"
"/options" 0 "options"
"/virtual-devices@100" 0 "vnex"
"/virtual-devices@100/channel-devices@200" 0 "cnex"
"/virtual-devices@100/channel-devices@200/disk@0" 0 "vdc"
"/virtual-devices@100/channel-devices@200/pciv-communication@0" 0 "vpci"
"/virtual-devices@100/channel-devices@200/network@0" 0 "vnet"
"/virtual-devices@100/channel-devices@200/network@1" 1 "vnet"
"/virtual-devices@100/channel-devices@200/network@2" 2 "vnet"
"/virtual-devices@100/channel-devices@200/network@3" 3 "vnet"
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc" << We want
this one
```

3. `/platform/sun4v/kernel/drv/vdc.conf`Editar :

```
block-size-list="1:4096";
```

Isso significa que a instância 1 do dispositivo recebe um tamanho de bloco de 4096MB.

Como exemplo adicional, suponha que as instâncias 1 a 6 do vdsk precisem ser configuradas para um tamanho de bloco 4K e `/etc/path_to_inst` lerem da seguinte forma:

```
"/virtual-devices@100/channel-devices@200/disk@1" 1 "vdc"
"/virtual-devices@100/channel-devices@200/disk@2" 2 "vdc"
"/virtual-devices@100/channel-devices@200/disk@3" 3 "vdc"
"/virtual-devices@100/channel-devices@200/disk@4" 4 "vdc"
"/virtual-devices@100/channel-devices@200/disk@5" 5 "vdc"
"/virtual-devices@100/channel-devices@200/disk@6" 6 "vdc"
```

4. O arquivo final `vdc.conf` deve conter o seguinte:

```
block-size-list="1:8192","2:8192","3:8192","4:8192","5:8192","6:8192";
```

> O LDOM deve ser reinicializado depois que o vdc.conf é configurado e o vdsk é criado. Este passo não pode ser evitado. A alteração do tamanho do bloco só entra em vigor após uma reinicialização. Prossiga com a configuração do zpool e certifique-se de que o ashift está corretamente configurado para 12, conforme descrito anteriormente.

Registo intenção ZFS (ZIL)

Geralmente, não há razão para localizar o ZFS Intent Log (ZIL) em um dispositivo diferente. O log pode compartilhar espaço com a piscina principal. O uso principal de um ZIL separado é quando se usa unidades físicas que não têm os recursos de armazenamento em cache de gravação em arrays de armazenamento modernos.

logbias

Defina `logbias` o parâmetro em sistemas de arquivos ZFS que hospedam dados Oracle.

```
zfs set logbias=throughput <filesystem>
```

O uso desse parâmetro reduz os níveis gerais de gravação. Sob os padrões, os dados escritos são comprometidos primeiro com o ZIL e depois para o pool de armazenamento principal. Essa abordagem é apropriada para uma configuração usando uma configuração de unidade simples, que inclui um dispositivo ZIL baseado em SSD e Mídia giratória para o pool de armazenamento principal. Isso ocorre porque permite que um commit ocorra em uma única transação de e/S na Mídia de menor latência disponível.

Ao usar um storage array moderno que inclua sua própria funcionalidade de armazenamento em cache, essa abordagem geralmente não é necessária. Em circunstâncias raras, pode ser desejável submeter uma gravação com uma única transação ao log, como uma carga de trabalho que consiste em gravações

aleatórias altamente concentradas e sensíveis à latência. Há consequências na forma de amplificação de gravação porque os dados registrados são gravados no pool de armazenamento principal, resultando em uma duplicação da atividade de gravação.

E/S direta

Muitas aplicações, incluindo produtos Oracle, podem ignorar o cache de buffer do host habilitando a e/S direta Esta estratégia não funciona como esperado com sistemas de arquivos ZFS. Embora o cache do buffer do host seja ignorado, o próprio ZFS continua a armazenar dados em cache. Essa ação pode resultar em resultados enganosos ao usar ferramentas como fio ou sio para realizar testes de desempenho, pois é difícil prever se a e/S está chegando ao sistema de armazenamento ou se está sendo armazenada em cache localmente no sistema operacional. Essa ação também torna muito difícil usar esses testes sintéticos para comparar o desempenho do ZFS com outros sistemas de arquivos. Na prática, há pouca ou nenhuma diferença no desempenho do sistema de arquivos em workloads reais do usuário.

Vários zpools

Backups, restaurações, clones e arquivamento baseados em snapshot de dados baseados em ZFS devem ser executados no nível do zpool e, geralmente, exigem vários zpools. Um zpool é análogo a um grupo de discos LVM e deve ser configurado usando as mesmas regras. Por exemplo, um banco de dados provavelmente é melhor definido com os datafiles que residem em `zpool1` e os logs de arquivo, arquivos de controle e logs de refazer que residem em `zpool2`. Essa abordagem permite um hot backup padrão no qual o banco de dados é colocado no modo hot backup, seguido por um snapshot de `zpool1`. O banco de dados é então removido do modo hot backup, o arquivo de log é forçado e um snapshot de `zpool2` é criado. Uma operação de restauração requer a desmontagem dos sistemas de arquivos zfs e a remoção do zpool em sua totalidade, seguindo-se uma operação de restauração do SnapRestore. O zpool pode então ser colocado on-line novamente e o banco de dados recuperado.

sistema de arquivos_options

O parâmetro Oracle `filesystemio_options` funciona de forma diferente com o ZFS. Se `setall` ou `directio` for usado, as operações de gravação são síncronas e ignoram o cache do buffer do sistema operacional, mas as leituras são armazenadas em buffer pelo ZFS. Essa ação causa dificuldades na análise de desempenho, pois às vezes, a e/S é intercetada e atendida pelo cache ZFS, tornando a latência do armazenamento e a e/S total menos do que parece ser.

Configuração de rede em sistemas AFF/ FAS

Interfaces lógicas

Os bancos de dados Oracle precisam de acesso ao armazenamento. Interfaces lógicas (LIFs) são o encanamento de rede que coneta uma máquina virtual de armazenamento (SVM) à rede e, por sua vez, ao banco de dados. O design de LIF adequado é necessário para garantir que haja largura de banda suficiente para cada workload de banco de dados, e o failover não resulta em perda de serviços de storage.

Esta seção fornece uma visão geral dos principais princípios de design de LIF. Para obter documentação mais abrangente, consulte o "Documentação do Gerenciamento de rede da ONTAP". Assim como em outros aspectos da arquitetura de banco de dados, as melhores opções para o design de máquina virtual de storage (SVM, conhecido como vserver na CLI) e interface lógica (LIF) dependem muito dos requisitos de dimensionamento e necessidades de negócios.

Considere os seguintes tópicos principais ao criar uma estratégia de LIF:

- **Desempenho.** A largura de banda da rede é suficiente?
- **Resiliência.** Há algum ponto único de falha no projeto?
- **Capacidade de gerenciamento.** A rede pode ser dimensionada sem interrupções?

Esses tópicos se aplicam à solução completa, desde o host até os switches até o sistema de storage.

Tipos de LIF

Existem vários tipos de LIF. "Documentação do ONTAP sobre tipos de LIF" Forneça informações mais completas sobre este tópico, mas de uma perspetiva funcional os LIFs podem ser divididos nos seguintes grupos:

- **LIFs de gerenciamento de clusters e nós.** LIFs usadas para gerenciar o cluster de armazenamento.
- **LIFs de gerenciamento de SVM.** Interfaces que permitem o acesso a um SVM por meio da API REST ou ONTAPI (também conhecida como ZAPI) para funções como criação de snapshot ou redimensionamento de volume. Produtos como o SnapManager para Oracle (SMO) precisam ter acesso a um LIF de gerenciamento de SVM.
- **LIFs de dados.** Interfaces para dados FC, iSCSI, NVMe/FC, NVMe/TCP, NFS ou SMB/CIFS.

> Um LIF de dados usado para tráfego NFS também pode ser usado para gerenciamento alterando a política de firewall de `data` para `mgmt` ou outra política que permita HTTP, HTTPS ou SSH. Essa alteração pode simplificar a configuração de rede evitando a configuração de cada host para acesso ao LIF de dados NFS e a um LIF de gerenciamento separado. Não é possível configurar uma interface para iSCSI e tráfego de gerenciamento, apesar do fato de ambos usarem um protocolo IP. Um LIF de gerenciamento separado é necessário em ambientes iSCSI.

Design de SAN LIF

O design de LIF em um ambiente SAN é relativamente simples por um motivo: Multipathing. Todas as implementações de SAN modernas permitem que um cliente acesse dados em vários caminhos de rede independentes e selecione o melhor caminho ou caminhos para acesso. Como resultado, o desempenho em relação ao design de LIF é mais simples de abordar, pois os clientes SAN equilibram automaticamente a e/S de carga nos melhores caminhos disponíveis.

Se um caminho ficar indisponível, o cliente selecionará automaticamente um caminho diferente. A simplicidade resultante do design torna os LIFs SAN geralmente mais gerenciáveis. Isso não significa que um ambiente SAN sempre seja gerenciado com mais facilidade, porque há muitos outros aspectos do storage SAN que são muito mais complicados do que o NFS. Isso significa simplesmente que o design de SAN LIF é mais fácil.

Desempenho

A consideração mais importante com o desempenho de LIF em um ambiente SAN é a largura de banda. Por exemplo, um cluster ONTAP AFF de dois nós com duas portas FC de 16GB GB por nó permite até 32GB Gbps de largura de banda de/para cada nó.

Resiliência

OS LIFs DE SAN não fazem failover em um sistema de storage da AFF. Se um LIF SAN falhar devido ao failover de controladora, o software multipathing do cliente deteta a perda de um caminho e redireciona e/S para um LIF diferente. Com os sistemas de armazenamento ASA, os LIFs serão falhados após um curto

atraso, mas isso não interrompe a e/S porque já existem caminhos ativos no outro controlador. O processo de failover ocorre para restaurar o acesso do host em todas as portas definidas.

Capacidade de gerenciamento

A migração de LIF é uma tarefa muito mais comum em um ambiente NFS, pois a migração de LIF geralmente é associada à realocação de volumes no cluster. Não é necessário migrar um LIF em um ambiente SAN quando os volumes são transferidos para o par de HA. Isso ocorre porque, após a conclusão da movimentação de volume, o ONTAP envia uma notificação à SAN sobre uma alteração nos caminhos e os clientes SAN reotimizam automaticamente. A migração de LIF com SAN está associada principalmente a grandes alterações de hardware físico. Por exemplo, se uma atualização sem interrupções dos controladores for necessária, um LIF SAN é migrado para o novo hardware. Se uma porta FC estiver defeituosa, um LIF pode ser migrado para uma porta não utilizada.

Recomendações de design

A NetApp faz as seguintes recomendações:

- Não crie mais caminhos do que o necessário. O número excessivo de caminhos torna o gerenciamento geral mais complicado e pode causar problemas com failover de caminho em alguns hosts. Além disso, alguns hosts têm limitações de caminho inesperadas para configurações como inicialização por SAN.
- Poucas configurações devem exigir mais de quatro caminhos para um LUN. O valor de ter mais de dois nós anunciando caminhos para LUNs é limitado porque o host agregado de um LUN fica inacessível se o nó proprietário do LUN e seu parceiro de HA falhar. A criação de caminhos em nós que não o par de HA principal não é útil em tal situação.
- Embora o número de caminhos de LUN visíveis possa ser gerenciado selecionando quais portas estão incluídas nas zonas FC, geralmente é mais fácil incluir todos os pontos de destino potenciais na zona FC e controlar a visibilidade de LUN no nível ONTAP.
- No ONTAP 8,3 e posterior, o recurso de mapeamento de LUN seletivo (SLM) é o padrão. Com o SLM, qualquer novo LUN é automaticamente anunciado a partir do nó que é proprietário do agregado subjacente e do parceiro de HA do nó. Esse arranjo evita a necessidade de criar conjuntos de portas ou configurar o zoneamento para limitar a acessibilidade de portas. Cada LUN está disponível no número mínimo de nós necessário para ter a melhor performance e resiliência. *Caso um LUN precise ser migrado para fora dos dois controladores, os nós adicionais podem ser adicionados com o `lun mapping add-reporting-nodes` comando para que os LUNs sejam anunciados nos novos nós. Isso cria caminhos SAN adicionais para os LUNs para migração de LUN. No entanto, o host deve executar uma operação de descoberta para usar os novos caminhos.
- Não se preocupe excessivamente com o tráfego indireto. É melhor evitar o tráfego indireto em um ambiente com uso intenso de e/S para o qual cada microssegundo de latência é crítico, mas o efeito de performance visível é insignificante para workloads típicos.

Design de LIF de NFS

Em contraste com os protocolos SAN, o NFS tem uma capacidade limitada de definir vários caminhos para os dados. As extensões NFS paralelas (pNFS) para NFSv4 abordam essa limitação, mas como as velocidades ethernet atingiram 100GBMbps e além raramente há valor na adição de caminhos adicionais.

Performance e resiliência

Embora a medição do desempenho do SAN LIF seja principalmente uma questão de calcular a largura de banda total de todos os caminhos principais, determinar o desempenho do NFS LIF requer uma visão mais detalhada da configuração exata da rede. Por exemplo, duas portas 10Gb podem ser configuradas como portas físicas brutas ou podem ser configuradas como um grupo de interfaces LACP (Link Aggregation Control

Protocol). Se eles estiverem configurados como um grupo de interfaces, várias políticas de balanceamento de carga estarão disponíveis que funcionam de forma diferente, dependendo se o tráfego é comutado ou roteado. Por fim, o Oracle Direct NFS (DNFS) oferece configurações de balanceamento de carga que não existem em nenhum cliente NFS do sistema operacional no momento.

Diferentemente dos protocolos SAN, os sistemas de arquivos NFS exigem resiliência na camada do protocolo. Por exemplo, um LUN é sempre configurado com multipathing habilitado, o que significa que vários canais redundantes estão disponíveis para o sistema de storage, cada um dos quais usa o protocolo FC. Um sistema de arquivos NFS, por outro lado, depende da disponibilidade de um único canal TCP/IP que só pode ser protegido na camada física. Esse arranjo é o motivo pelo qual existem opções como failover de portas e agregação de portas LACP.

Em um ambiente NFS, a performance e a resiliência são fornecidas na camada de protocolo de rede. Como resultado, ambos os tópicos estão interligados e devem ser discutidos juntos.

Vincular LIFs a grupos de portas

Para vincular um LIF a um grupo de portas, associe o endereço IP de LIF a um grupo de portas físicas. O método principal para agregar portas físicas em conjunto é o LACP. A capacidade de tolerância a falhas do LACP é bastante simples; cada porta em um grupo LACP é monitorada e removida do grupo de portas em caso de mau funcionamento. Existem, no entanto, muitos equívocos sobre como o LACP funciona em relação ao desempenho:

- O LACP não requer a configuração no switch para corresponder ao endpoint. Por exemplo, o ONTAP pode ser configurado com balanceamento de carga baseado em IP, enquanto um switch pode usar balanceamento de carga baseado em MAC.
- Cada endpoint que usa uma conexão LACP pode escolher independentemente a porta de transmissão de pacotes, mas não pode escolher a porta usada para recebimento. Isso significa que o tráfego de ONTAP para um destino específico está vinculado a uma porta específica, e o tráfego de retorno pode chegar em uma interface diferente. No entanto, isso não causa problemas.
- O LACP não distribui uniformemente o tráfego o tempo todo. Em um ambiente grande com muitos clientes NFS, o resultado é normalmente até mesmo o uso de todas as portas em uma agregação LACP. No entanto, qualquer sistema de arquivos NFS no ambiente é limitado à largura de banda de apenas uma porta, e não a agregação inteira.
- Embora as políticas LACP robin estejam disponíveis no ONTAP, essas políticas não abordam a conexão de um switch para um host. Por exemplo, uma configuração com um tronco LACP de quatro portas em um host e um tronco LACP de quatro portas no ONTAP ainda é capaz de ler apenas um sistema de arquivos usando uma única porta. Embora o ONTAP possa transmitir dados através das quatro portas, não há tecnologias de switch disponíveis que sejam enviadas do switch para o host através das quatro portas. Apenas um é usado.

A abordagem mais comum em ambientes maiores que consistem em muitos hosts de banco de dados é construir um agregado LACP de um número apropriado de interfaces 10Gb (ou mais rápidas) usando o balanceamento de carga IP. Essa abordagem permite que o ONTAP forneça uso uniforme de todas as portas, desde que existam clientes suficientes. O balanceamento de carga é interrompido quando há menos clientes na configuração porque o entroncamento LACP não redistribui dinamicamente a carga.

Quando uma conexão é estabelecida, o tráfego em uma determinada direção é colocado em apenas uma porta. Por exemplo, um banco de dados que executa uma verificação de tabela completa em um sistema de arquivos NFS conetado por meio de um tronco LACP de quatro portas lê dados através de apenas uma placa de interface de rede (NIC). Se apenas três servidores de banco de dados estiverem em tal ambiente, é possível que todos os três estejam lendo da mesma porta, enquanto as outras três portas estiverem ociosas.

Vincule LIFs a portas físicas

Vincular um LIF a uma porta física resulta em um controle mais granular sobre a configuração de rede, pois um determinado endereço IP em um sistema ONTAP está associado a apenas uma porta de rede de cada vez. A resiliência é então realizada por meio da configuração de grupos de failover e políticas de failover.

Políticas de failover e grupos de failover

O comportamento dos LIFs durante a interrupção da rede é controlado por políticas de failover e grupos de failover. As opções de configuração foram alteradas com as diferentes versões do ONTAP. Consulte o "Documentação de gerenciamento de rede ONTAP para grupos e políticas de failover" para obter detalhes específicos sobre a versão do ONTAP que está sendo implantado.

O ONTAP 8,3 e superior permitem o gerenciamento de failover de LIF com base em domínios de broadcast. Portanto, um administrador pode definir todas as portas que têm acesso a uma determinada sub-rede e permitir que o ONTAP selecione um LIF de failover apropriado. Essa abordagem pode ser usada por alguns clientes, mas tem limitações em um ambiente de rede de storage de alta velocidade devido à falta de previsibilidade. Por exemplo, um ambiente pode incluir ambas as portas 1GB para acesso de rotina ao sistema de arquivos e portas 10Gb para e/S de arquivo de dados Se ambos os tipos de portas existirem no mesmo domínio de broadcast, o failover de LIF pode resultar na movimentação de e/S de um arquivo de dados de uma porta 10Gb para uma porta 1GB.

Em resumo, considere as seguintes práticas:

1. Configurar um grupo de failover conforme definido pelo usuário.
2. Preencha o grupo de failover com portas na controladora de parceiro de failover de storage (SFO) para que as LIFs sigam os agregados durante um failover de storage. Isso evita a criação de tráfego indireto.
3. Use portas de failover com caraterísticas de desempenho correspondentes ao LIF original. Por exemplo, um LIF em uma única porta 10Gb física deve incluir um grupo de failover com uma única porta 10Gb. Um LIF LACP de quatro portas deve falhar para outro LIF LACP de quatro portas. Essas portas seriam um subconjunto das portas definidas no domínio de broadcast.
4. Defina a política de failover como somente parceiro SFO. Isso garante que o LIF siga o agregado durante o failover.

Reversão automática

Defina `auto-revert` o parâmetro conforme desejado. A maioria dos clientes prefere definir este parâmetro para que `true` o LIF reverta para sua porta inicial. No entanto, em alguns casos, os clientes definiram isso como "falso" que um failover inesperado pode ser investigado antes de retornar um LIF à sua porta inicial.

Relação LIF-volume

Um equívoco comum é que deve haver uma relação do 1:1 entre volumes e LIFs NFS. Embora essa configuração seja necessária para mover um volume em qualquer lugar em um cluster, sem nunca criar tráfego de interconexão adicional, ela não é categoricamente um requisito. O tráfego entre clusters deve ser considerado, mas a mera presença de tráfego entre clusters não cria problemas. Muitos dos benchmarks publicados criados para o ONTAP incluem predominantemente I/O. indireto

Por exemplo, um projeto de banco de dados contendo um número relativamente pequeno de bancos de dados críticos ao desempenho que exigiam apenas um total de 40 volumes pode garantir um volume 1:1 para a estratégia LIF, um arranjo que exigiria 40 endereços IP. Qualquer volume poderia então ser movido para qualquer lugar do cluster junto com o LIF associado, e o tráfego sempre seria direto, minimizando cada fonte de latência, mesmo nos níveis de microssegundos.

Como um exemplo de contador, um ambiente grande e hospedado pode ser mais facilmente gerenciado com um relacionamento 1:1 entre clientes e LIFs. Com o tempo, um volume pode precisar ser migrado para um nó diferente, o que causaria algum tráfego indireto. No entanto, o efeito de desempenho deve ser indetetável, a menos que as portas de rede no switch de interconexão estejam saturando. Se houver problema, um novo LIF pode ser estabelecido em nós adicionais e o host pode ser atualizado na próxima janela de manutenção para remover o tráfego indireto da configuração.

Configuração TCP/IP e ethernet

Muitos clientes do Oracle no ONTAP usam ethernet, o protocolo de rede de NFS, iSCSI, NVMe/TCP e, especialmente, a nuvem.

Configurações do sistema operacional do host

A maioria da documentação do fornecedor de aplicativos inclui configurações específicas de TCP e ethernet destinadas a garantir que o aplicativo esteja funcionando de forma ideal. Essas mesmas configurações geralmente são suficientes para oferecer também um desempenho ideal de storage baseado em IP.

Controle de fluxo Ethernet

Esta tecnologia permite que um cliente solicite que um remetente pare temporariamente a transmissão de dados. Isso geralmente é feito porque o recetor não consegue processar os dados de entrada com rapidez suficiente. Ao mesmo tempo, solicitar que um remetente cessasse a transmissão era menos disruptivo do que ter um recetor descartar pacotes porque os buffers estavam cheios. Este não é mais o caso com as pilhas TCP usadas nos sistemas operacionais atuais. Na verdade, o controle de fluxo causa mais problemas do que resolve.

Os problemas de desempenho causados pelo controle de fluxo Ethernet têm aumentado nos últimos anos. Isso ocorre porque o controle de fluxo Ethernet opera na camada física. Se uma configuração de rede permitir que qualquer sistema operacional host envie uma solicitação de controle de fluxo Ethernet para um sistema de armazenamento, o resultado será uma pausa em e/S para todos os clientes conetados. Como um número crescente de clientes é servido por um único controlador de armazenamento, a probabilidade de um ou mais desses clientes enviar solicitações de controle de fluxo aumenta. O problema tem sido visto frequentemente em sites de clientes com ampla virtualização de SO.

Uma NIC em um sistema NetApp não deve receber solicitações de controle de fluxo. O método utilizado para alcançar este resultado varia consoante o fabricante do comutador de rede. Na maioria dos casos, o controle de fluxo em um switch Ethernet pode ser definido como `receive desired` `receive on` ou , o que significa que uma solicitação de controle de fluxo não é encaminhada para o controlador de armazenamento. Em outros casos, a conexão de rede no controlador de armazenamento pode não permitir a desativação do controle de fluxo. Nesses casos, os clientes devem ser configurados para nunca enviar solicitações de controle de fluxo, seja alterando para a configuração da NIC no próprio servidor host ou para as portas de switch às quais o servidor host está conetado.

A NetApp recomenda certificando-se de que os controladores de armazenamento NetApp não recebam pacotes de controle de fluxo Ethernet. Isso geralmente pode ser feito definindo as portas do switch às quais o controlador está conetado, mas alguns hardwares do switch têm limitações que podem exigir alterações no lado do cliente.

Tamanhos de MTU

O uso de quadros jumbo tem sido mostrado para oferecer alguma melhoria de desempenho em redes 1GBG, reduzindo a sobrecarga de CPU e rede, mas o benefício geralmente não é significativo.

A NetApp recomenda a implementação de quadros jumbo quando possível, tanto para obter quaisquer benefícios potenciais de desempenho quanto para preparar a solução para o futuro.

Usar quadros jumbo em uma rede 10GbG é quase obrigatório. Isso ocorre porque a maioria das implementações 10Gb atingem um limite de pacotes por segundo sem quadros jumbo antes que atinjam a marca 10Gb. O uso de quadros jumbo melhora a eficiência no processamento TCP/IP porque permite que o sistema operacional, servidor, NICs e o sistema de armazenamento processem menos pacotes, mas maiores. A melhoria do desempenho varia de NIC para NIC, mas é significativa.

Para implementações de quadros jumbo, existe a crença comum, mas incorreta, de que todos os dispositivos conetados devem suportar quadros jumbo e que o tamanho da MTU deve corresponder de ponta a ponta Em vez disso, os dois pontos de extremidade da rede negociam o tamanho de quadro mais alto mutuamente aceitável ao estabelecer uma conexão. Em um ambiente típico, um switch de rede é definido para um tamanho MTU de 9216, o controlador NetApp é definido como 9000 e os clientes são definidos como uma combinação de 9000 e 1514. Os clientes que podem suportar um MTU de 9000 podem usar quadros jumbo, e os clientes que só podem suportar 1514 podem negociar um valor mais baixo.

Problemas com este arranjo são raros em um ambiente completamente comutado. No entanto, tenha cuidado em um ambiente roteado que nenhum roteador intermediário é forçado a fragmentar quadros jumbo.

A NetApp recomenda configurando o seguinte:

- Os quadros jumbo são desejáveis, mas não são necessários com Ethernet de 1GB GB (GbE).
- Os quadros Jumbo são necessários para o máximo desempenho com 10GbE e mais rápido.

Parâmetros TCP

Três configurações geralmente são mal configuradas: Carimbos de data/hora TCP, reconhecimento seletivo (SACK) e escala de janela TCP. Muitos documentos desatualizados na Internet recomendam desativar um ou mais desses parâmetros para melhorar o desempenho. Havia algum mérito a esta recomendação há muitos anos, quando os recursos da CPU eram muito menores e havia um benefício para reduzir a sobrecarga no processamento TCP sempre que possível.

No entanto, com os sistemas operacionais modernos, desabilitar qualquer um desses recursos TCP geralmente resulta em nenhum benefício detetável, ao mesmo tempo em que potencialmente danifica o desempenho. Os danos no desempenho são especialmente prováveis em ambientes de rede virtualizados, pois esses recursos são necessários para o manuseio eficiente da perda de pacotes e alterações na qualidade da rede.

A NetApp recomenda a ativação de timestamps TCP, SACK e escala de janelas TCP no host, e todos esses três parâmetros devem estar ativados por padrão em qualquer sistema operacional atual.

Configuração de FC SAN

A configuração do FC SAN para bancos de dados Oracle trata principalmente de seguir as práticas recomendadas diárias de SAN.

Isso inclui medidas de Planejamento típicas, como garantir a existência de largura de banda suficiente na SAN entre o host e o sistema de storage, verificar se todos os caminhos de SAN existem entre todos os

dispositivos necessários, usar as configurações de porta FC exigidas pelo fornecedor do switch FC, evitar a contenção de ISL e usar o monitoramento adequado da malha SAN.

Zoneamento

Uma zona FC nunca deve conter mais de um iniciador. Tal arranjo pode parecer funcionar inicialmente, mas a conversa cruzada entre iniciadores eventualmente interfere com o desempenho e a estabilidade.

As zonas de destino múltiplo são geralmente consideradas seguras, embora em raras circunstâncias o comportamento das portas de destino FC de diferentes fornecedores tenha causado problemas. Por exemplo, evite incluir as portas de destino de um storage array NetApp e não NetApp na mesma zona. Além disso, é ainda mais provável que colocar um sistema de storage NetApp e um dispositivo de fita na mesma zona cause problemas.

Rede de conexão direta

Às vezes, os administradores de storage preferem simplificar suas infraestruturas removendo switches de rede da configuração. Isso pode ser suportado em alguns cenários.

ISCSI e NVMe/TCP

Um host usando iSCSI ou NVMe/TCP pode ser conetado diretamente a um sistema de storage e operar normalmente. A razão é pathing. Conexões diretas a dois controladores de storage diferentes resultam em dois caminhos independentes para o fluxo de dados. A perda de caminho, porta ou controlador não impede que o outro caminho seja usado.

NFS

O armazenamento NFS com conexão direta pode ser usado, mas com uma limitação significativa - o failover não funcionará sem um esforço significativo de script, o que seria da responsabilidade do cliente.

O motivo pelo qual o failover sem interrupções é complicado com o storage NFS com conexão direta é o roteamento que ocorre no sistema operacional local. Por exemplo, suponha que um host tenha um endereço IP de 192.168.1.1/24 e esteja conetado diretamente a um controlador ONTAP com um endereço IP de 192.168.1.50/24. Durante o failover, esse endereço 192.168.1.50 pode fazer failover para a outra controladora e estará disponível para o host, mas como o host deteta sua presença? O endereço 192.168.1.1 original ainda existe na NIC host que não se coneta mais a um sistema operacional. O tráfego destinado a 192.168.1.50 continuaria a ser enviado para uma porta de rede inoperável.

A segunda NIC do SO poderia ser configurada como 19 2.168.1.2 e seria capaz de se comunicar com o endereço 192.168.1.50 com falha, mas as tabelas de roteamento local teriam um padrão de usar um endereço **e apenas um** para se comunicar com a sub-rede 192.168.1.0/24. Um sysadmin poderia criar uma estrutura de script que detetaria uma conexão de rede com falha e alteraria as tabelas de roteamento local ou colocaria interfaces para cima e para baixo. O procedimento exato dependeria do SO em uso.

Na prática, os clientes da NetApp têm NFS com conexão direta, mas normalmente apenas para workloads em que as pausas de e/S durante failovers são aceitáveis. Quando os suportes rígidos são usados, não deve haver nenhum erro de e/S durante essas pausas. O IO deve travar até que os serviços sejam restaurados, seja por um failback ou intervenção manual para mover endereços IP entre NICs no host.

Conexão direta FC

Não é possível conectar diretamente um host a um sistema de storage ONTAP usando o protocolo FC. A razão é o uso de NPIV. A WWN que identifica uma porta ONTAP FC para a rede FC usa um tipo de virtualização chamado NPIV. Qualquer dispositivo conetado a um sistema ONTAP deve ser capaz de reconhecer um NPIV WWN. Não há fornecedores atuais de HBA que ofereçam um HBA que possa ser instalado em um host que possa suportar um destino NPIV.

Configuração de rede em sistemas ASA r2

Interfaces lógicas

Os bancos de dados Oracle precisam de acesso ao armazenamento. Interfaces lógicas (LIFs) são o encanamento de rede que coneta uma máquina virtual de armazenamento (SVM) à rede e, por sua vez, ao banco de dados. O design de LIF adequado é necessário para garantir que haja largura de banda suficiente para cada workload de banco de dados, e o failover não resulta em perda de serviços de storage.

Esta seção fornece uma visão geral dos principais princípios de design LIF para sistemas ASA r2, otimizados para ambientes exclusivamente SAN. Para obter documentação mais completa, consulte o "Documentação do Gerenciamento de rede da ONTAP". Assim como em outros aspectos da arquitetura de banco de dados, as melhores opções para o projeto de máquina virtual de armazenamento (SVM, conhecida como vserver na linha de comando) e interface lógica (LIF) dependem muito dos requisitos de escalabilidade e das necessidades do negócio.

Considere os seguintes tópicos principais ao criar uma estratégia de LIF:

- **Desempenho.** A largura de banda da rede é suficiente para as cargas de trabalho do Oracle?
- **Resiliência.** Há algum ponto único de falha no projeto?
- **Capacidade de gerenciamento.** A rede pode ser dimensionada sem interrupções?

Esses tópicos se aplicam à solução completa, desde o host até os switches até o sistema de storage.

Tipos de LIF

Existem vários tipos de LIF. "Documentação do ONTAP sobre tipos de LIF" Forneça informações mais completas sobre este tópico, mas de uma perspetiva funcional os LIFs podem ser divididos nos seguintes grupos:

- **LIFs de gerenciamento de clusters e nós.** LIFs usadas para gerenciar o cluster de armazenamento.
- **LIFs de gerenciamento de SVM.** Interfaces que permitem o acesso a um SVM por meio da API REST ou ONTAPI (também conhecida como ZAPI) para funções como criação de snapshot ou redimensionamento de volume. Produtos como o SnapManager para Oracle (SMO) precisam ter acesso a um LIF de gerenciamento de SVM.
- **Fichas de dados LIF.** Interfaces compatíveis apenas com protocolos SAN: FC, iSCSI, NVMe/FC, NVMe/TCP. Os protocolos NAS (NFS, SMB/CIFS) não são suportados em sistemas ASA r2.

Não é possível configurar uma interface para tráfego iSCSI (ou NVMe/TCP) e tráfego de gerenciamento simultaneamente, apesar de ambos utilizarem um protocolo IP. É necessário um LIF de gerenciamento separado em ambientes iSCSI ou NVMe/TCP. Para garantir resiliência e desempenho, configure várias LIFs de dados SAN por protocolo por nó e distribua-as por diferentes portas físicas e estruturas. Diferentemente dos sistemas AFF/ FAS , o ASA r2 não permite tráfego NFS ou SMB, portanto não há opção de reutilizar uma LIF de dados NAS para gerenciamento.

Design de SAN LIF

O design de LIF em um ambiente SAN é relativamente simples por um motivo: Multipathing. Todas as implementações de SAN modernas permitem que um cliente acesse dados em vários caminhos de rede independentes e selecione o melhor caminho ou caminhos para acesso. Como resultado, o desempenho em relação ao design de LIF é mais simples de abordar, pois os clientes SAN equilibram automaticamente a e/S de carga nos melhores caminhos disponíveis.

Se um caminho ficar indisponível, o cliente selecionará automaticamente um caminho diferente. A simplicidade resultante do design torna os LIFs SAN geralmente mais gerenciáveis. Isso não significa que um ambiente SAN sempre seja gerenciado com mais facilidade, porque há muitos outros aspectos do storage SAN que são muito mais complicados do que o NFS. Isso significa simplesmente que o design de SAN LIF é mais fácil.

Desempenho

O fator mais importante a considerar em relação ao desempenho de uma LIF em um ambiente SAN é a largura de banda. Por exemplo, um cluster ASA r2 de dois nós com duas portas FC de 32 Gb por nó permite até 64 Gb de largura de banda de/para cada nó. Da mesma forma, para NVMe/TCP ou iSCSI, assegure conectividade suficiente de 25GbE ou 100GbE para cargas de trabalho Oracle.

Resiliência

As interfaces LIF SAN não realizam failover da mesma forma que as interfaces LIF NAS. Os sistemas ASA r2 dependem de multipathing do host (MPIO/ALUA) para resiliência. Se uma SAN LIF ficar indisponível devido a uma falha do controlador, o software de multipathing do cliente detecta a perda de um caminho e redireciona a E/S para um caminho alternativo. O ASA r2 pode realizar a realocação do LIF após um breve atraso para restaurar a disponibilidade total do caminho, mas isso não interrompe a E/S porque já existem caminhos ativos no nó parceiro. O processo de failover ocorre para restaurar o acesso do host em todas as portas definidas.

Capacidade de gerenciamento

Não há necessidade de migrar uma LIF em um ambiente SAN quando os volumes são realocados dentro do par HA. Isso ocorre porque, após a conclusão da movimentação do volume, o ONTAP envia uma notificação ao SAN sobre uma alteração nos caminhos, e os clientes SAN se reotimizam automaticamente. A migração de LIF com SAN está principalmente associada a grandes alterações físicas de hardware. Por exemplo, se for necessária uma atualização não disruptiva dos controladores, uma SAN LIF é migrada para o novo hardware. Caso uma porta FC apresente defeito, uma LIF pode ser migrada para uma porta não utilizada.

Recomendações de design

A NetApp faz as seguintes recomendações para ambientes SAN ASA r2:

- Não crie mais caminhos do que o necessário. O número excessivo de caminhos torna o gerenciamento geral mais complicado e pode causar problemas com failover de caminho em alguns hosts. Além disso,

alguns hosts têm limitações de caminho inesperadas para configurações como inicialização por SAN.

- Poucas configurações devem exigir mais de quatro caminhos para um LUN. O valor de ter mais de dois nós anunciando caminhos para LUNs é limitado porque o host agregado de um LUN fica inacessível se o nó proprietário do LUN e seu parceiro de HA falhar. A criação de caminhos em nós que não o par de HA principal não é útil em tal situação.
- Embora o número de caminhos de LUN visíveis possa ser gerenciado selecionando quais portas estão incluídas nas zonas FC, geralmente é mais fácil incluir todos os pontos de destino potenciais na zona FC e controlar a visibilidade de LUN no nível ONTAP.
- Utilize o recurso de mapeamento seletivo de LUN (SLM), que está ativado por padrão. Com o SLM, qualquer novo LUN é automaticamente anunciado a partir do nó que possui o agregado subjacente e do parceiro HA do nó. Essa configuração evita a necessidade de criar conjuntos de portas ou configurar o zoneamento para limitar o acesso às portas. Cada LUN está disponível no número mínimo de nós necessários para desempenho e resiliência ideais.
- Caso seja necessário migrar um LUN para fora dos dois controladores, os nós adicionais podem ser adicionados com o `lun mapping add-reporting-nodes` comando para que os LUNs sejam anunciados nos novos nós. Ao fazer isso, são criados caminhos SAN adicionais para os LUNs, permitindo a migração dos mesmos. No entanto, o host precisa realizar uma operação de descoberta para usar os novos caminhos.
- Não se preocupe excessivamente com o tráfego indireto. É melhor evitar o tráfego indireto em um ambiente com uso intenso de e/S para o qual cada microssegundo de latência é crítico, mas o efeito de performance visível é insignificante para workloads típicos.

Configuração TCP/IP e ethernet

Muitos clientes do Oracle no ASA r2 ONTAP usam Ethernet, o protocolo de rede do iSCSI e NVMe/TCP.

Configurações do sistema operacional do host

A maioria da documentação do fornecedor de aplicativos inclui configurações específicas de TCP e ethernet destinadas a garantir que o aplicativo esteja funcionando de forma ideal. Essas mesmas configurações geralmente são suficientes para oferecer também um desempenho ideal de storage baseado em IP.

Controle de fluxo Ethernet

Esta tecnologia permite que um cliente solicite que um remetente pare temporariamente a transmissão de dados. Isso geralmente é feito porque o recetor não consegue processar os dados de entrada com rapidez suficiente. Ao mesmo tempo, solicitar que um remetente cessasse a transmissão era menos disruptivo do que ter um recetor descartar pacotes porque os buffers estavam cheios. Este não é mais o caso com as pilhas TCP usadas nos sistemas operacionais atuais. Na verdade, o controle de fluxo causa mais problemas do que resolve.

Os problemas de desempenho causados pelo controle de fluxo Ethernet têm aumentado nos últimos anos. Isso ocorre porque o controle de fluxo Ethernet opera na camada física. Se uma configuração de rede permitir que qualquer sistema operacional host envie uma solicitação de controle de fluxo Ethernet para um sistema de armazenamento, o resultado será uma pausa em e/S para todos os clientes conetados. Como um número crescente de clientes é servido por um único controlador de armazenamento, a probabilidade de um ou mais desses clientes enviar solicitações de controle de fluxo aumenta. O problema tem sido visto frequentemente em sites de clientes com ampla virtualização de SO.

Uma NIC em um sistema NetApp não deve receber solicitações de controle de fluxo. O método utilizado para alcançar este resultado varia consoante o fabricante do comutador de rede. Na maioria dos casos, o controle

de fluxo em um switch Ethernet pode ser definido como `receive desired` `receive on` ou , o que significa que uma solicitação de controle de fluxo não é encaminhada para o controlador de armazenamento. Em outros casos, a conexão de rede no controlador de armazenamento pode não permitir a desativação do controle de fluxo. Nesses casos, os clientes devem ser configurados para nunca enviar solicitações de controle de fluxo, seja alterando para a configuração da NIC no próprio servidor host ou para as portas de switch às quais o servidor host está conetado.

Para sistemas ASA r2, que são exclusivamente SAN, as considerações de controle de fluxo Ethernet aplicam-se principalmente ao tráfego iSCSI e NVMe/TCP.

* A NetApp recomenda* garantir que os controladores de armazenamento NetApp ASA r2 não recebam pacotes de controle de fluxo Ethernet. Isso geralmente pode ser feito configurando as portas do switch às quais o controlador está conectado, mas alguns switches têm limitações que podem exigir alterações no lado do cliente.

Tamanhos de MTU

O uso de quadros jumbo tem sido mostrado para oferecer alguma melhoria de desempenho em redes 1GBG, reduzindo a sobrecarga de CPU e rede, mas o benefício geralmente não é significativo.

A NetApp recomenda a implementação de quadros jumbo quando possível, tanto para obter quaisquer benefícios potenciais de desempenho quanto para preparar a solução para o futuro.

Para sistemas ASA r2, que são exclusivamente SAN, os jumbo frames aplicam-se apenas a protocolos SAN baseados em Ethernet (iSCSI e NVMe/TCP).

Usar quadros jumbo em uma rede 10GbG é quase obrigatório. Isso ocorre porque a maioria das implementações 10Gb atingem um limite de pacotes por segundo sem quadros jumbo antes que atinjam a marca 10Gb. O uso de quadros jumbo melhora a eficiência no processamento TCP/IP porque permite que o sistema operacional, servidor, NICs e o sistema de armazenamento processem menos pacotes, mas maiores. A melhoria do desempenho varia de NIC para NIC, mas é significativa.

Para implementações de quadros jumbo, existe a crença comum, mas incorreta, de que todos os dispositivos conetados devem suportar quadros jumbo e que o tamanho da MTU deve corresponder de ponta a ponta Em vez disso, os dois pontos de extremidade da rede negociam o tamanho de quadro mais alto mutuamente aceitável ao estabelecer uma conexão. Em um ambiente típico, um switch de rede é definido para um tamanho MTU de 9216, o controlador NetApp é definido como 9000 e os clientes são definidos como uma combinação de 9000 e 1514. Os clientes que podem suportar um MTU de 9000 podem usar quadros jumbo, e os clientes que só podem suportar 1514 podem negociar um valor mais baixo.

Problemas com este arranjo são raros em um ambiente completamente comutado. No entanto, tenha cuidado em um ambiente roteado que nenhum roteador intermediário é forçado a fragmentar quadros jumbo.

- A NetApp recomenda* configurar o seguinte para ambientes SAN ASA r2:
- Quadros jumbo são desejáveis, mas não obrigatórios com 1GbE.
- Para obter o máximo desempenho com 10GbE e velocidades superiores para tráfego iSCSI e NVMe/TCP, são necessários jumbo frames.

Parâmetros TCP

Três configurações geralmente são mal configuradas: Carimbos de data/hora TCP, reconhecimento seletivo (SACK) e escala de janela TCP. Muitos documentos desatualizados na Internet recomendam desativar um ou

mais desses parâmetros para melhorar o desempenho. Havia algum mérito a esta recomendação há muitos anos, quando os recursos da CPU eram muito menores e havia um benefício para reduzir a sobrecarga no processamento TCP sempre que possível.

No entanto, com os sistemas operacionais modernos, desabilitar qualquer um desses recursos TCP geralmente resulta em nenhum benefício detetável, ao mesmo tempo em que potencialmente danifica o desempenho. Os danos no desempenho são especialmente prováveis em ambientes de rede virtualizados, pois esses recursos são necessários para o manuseio eficiente da perda de pacotes e alterações na qualidade da rede.

A NetApp recomenda a ativação de timestamps TCP, SACK e escala de janelas TCP no host, e todos esses três parâmetros devem estar ativados por padrão em qualquer sistema operacional atual.

Configuração de FC SAN

A configuração de FC SAN para bancos de dados Oracle em sistemas ASA r2 consiste principalmente em seguir as melhores práticas padrão de SAN.

O ASA r2 é otimizado para cargas de trabalho exclusivas de SAN, portanto, os princípios permanecem os mesmos do AFF/ FAS, com foco em desempenho, resiliência e simplicidade. Isso inclui medidas típicas de planejamento, como garantir largura de banda suficiente na SAN entre o host e o sistema de armazenamento, verificar se todos os caminhos SAN existem entre todos os dispositivos necessários, usar as configurações de porta FC exigidas pelo fornecedor do switch FC, evitar conflitos de ISL e usar o monitoramento adequado da estrutura SAN.

Zoneamento

Uma zona FC nunca deve conter mais de um iniciador. Tal arranjo pode parecer funcionar inicialmente, mas a conversa cruzada entre iniciadores eventualmente interfere com o desempenho e a estabilidade.

As zonas de destino múltiplo são geralmente consideradas seguras, embora em raras circunstâncias o comportamento das portas de destino FC de diferentes fornecedores tenha causado problemas. Por exemplo, evite incluir as portas de destino de um storage array NetApp e não NetApp na mesma zona. Além disso, é ainda mais provável que colocar um sistema de storage NetApp e um dispositivo de fita na mesma zona cause problemas.

- O ASA r2 usa Zonas de Disponibilidade de Armazenamento em vez de agregados, mas isso não altera os princípios de zoneamento do FC.
- O multipathing (MPIO) continua sendo o principal mecanismo de resiliência; no entanto, para sistemas ASA r2 que suportam multipathing ativo-ativo simétrico, todos os caminhos para um LUN estão ativos e são usados para E/S simultaneamente.

Rede de conexão direta

Às vezes, os administradores de storage preferem simplificar suas infraestruturas removendo switches de rede da configuração. Isso pode ser suportado em alguns cenários.

ISCSI e NVMe/TCP

Um host que utiliza iSCSI ou NVMe/TCP pode ser conectado diretamente a um sistema de armazenamento ASA r2 e operar normalmente. O motivo é o roteamento. A conexão direta a dois controladores de armazenamento diferentes resulta em dois caminhos independentes para o fluxo de dados. A perda de um caminho, porta ou controlador não impede que o outro caminho seja usado, desde que o multipathing esteja configurado corretamente.

Conexão direta FC

Não é possível conectar diretamente um host a um sistema de armazenamento ASA r2 usando o protocolo FC. O motivo é o mesmo dos sistemas AFF/ FAS , ou seja, o uso de NPIV. O WWN que identifica uma porta ONTAP FC para a rede FC usa um tipo de virtualização chamado NPIV. Qualquer dispositivo conectado a um sistema ONTAP deve ser capaz de reconhecer um WWN NPIV. Atualmente, não existem fornecedores de HBA que ofereçam um HBA que possa ser instalado em um host capaz de suportar um alvo NPIV.

Configuração de storage em sistemas AFF/FAS

FC SAN

Alinhamento LUN

Alinhamento LUN refere-se a otimizar e/S em relação ao layout do sistema de arquivos subjacente.

Em um sistema ONTAP, o storage é organizado em 4KB unidades. Um bloco 8KB do banco de dados ou do sistema de arquivos deve ser mapeado para exatamente dois blocos 4KB. Se um erro na configuração de LUN mudar o alinhamento em 1KB em qualquer direção, cada bloco 8KB existiria em três blocos de armazenamento 4KB diferentes em vez de dois. Esse arranjo causaria maior latência e causaria a realização de e/S adicionais no sistema de storage.

O alinhamento também afeta arquiteturas LVM. Se um volume físico dentro de um grupo de volumes lógicos for definido em todo o dispositivo da unidade (nenhuma partição é criada), o primeiro bloco 4KB no LUN se alinha com o primeiro bloco 4KB no sistema de armazenamento. Este é um alinhamento correto. Problemas surgem com partições porque eles mudam o local inicial onde o sistema operacional usa o LUN. Desde que o deslocamento seja deslocado em unidades inteiras de 4KB, o LUN é alinhado.

Em ambientes Linux, crie grupos de volume lógicos em todo o dispositivo de unidade. Quando uma partição for necessária, verifique o alinhamento executando `fdisk -u` e verificando se o início de cada partição é um múltiplo de oito. Isso significa que a partição começa em um múltiplo de oito setores de 512 bytes, que é 4KB.

Consulte também a discussão sobre o alinhamento do bloco de compressão na "Eficiência"seção . Qualquer layout que esteja alinhado com os limites do bloco de compressão 8KBD também está alinhado com os limites 4KBD.

Avisos de desalinhamento

O log de refazer/transações do banco de dados normalmente gera e/S desalinhadas que podem causar avisos enganosos sobre LUNs desalinhados no ONTAP.

O log executa uma gravação sequencial do arquivo de log com gravações de tamanho variável. Uma operação de gravação de log que não esteja alinhada aos limites do 4KB normalmente não causa problemas de desempenho porque a próxima operação de gravação de log completa o bloco. O resultado é que o ONTAP é capaz de processar quase todas as gravações como blocos 4KB completos, mesmo que os dados

em cerca de 4KB blocos tenham sido gravados em duas operações separadas.

Verifique o alinhamento usando utilitários como `sio` ou `dd` que podem gerar e/S em um tamanho de bloco definido. As estatísticas de alinhamento de e/S no sistema de storage podem ser visualizadas com o `stats` comando. Consulte "Verificação do alinhamento do WAFL" para obter mais informações.

O alinhamento em ambientes Solaris é mais complicado. "Configuração do host SAN ONTAP"Consulte para obter mais informações.

Cuidado
Nos ambientes Solaris x86, tenha cuidado adicional com o alinhamento adequado, pois a maioria das configurações tem várias camadas de partições. Os cortes de partição do Solaris x86 geralmente existem em cima de uma tabela de partição de Registro de inicialização principal padrão.

Dimensionamento de LUN e contagem de LUN

Selecionar o tamanho ideal de LUN e o número de LUNs a serem usados é essencial para obter o desempenho e a capacidade de gerenciamento ideais dos bancos de dados Oracle.

Um LUN é um objeto virtualizado no ONTAP que existe em todas as unidades no agregado de hospedagem. Como resultado, o desempenho do LUN não é afetado pelo seu tamanho porque o LUN se baseia no potencial de desempenho total do agregado, independentemente do tamanho escolhido.

Por uma questão de conveniência, os clientes podem querer usar um LUN de um tamanho específico. Por exemplo, se um banco de dados for construído em um grupo de discos LVM ou Oracle ASM composto por dois LUNs de 1TB cada, esse grupo de discos deve ser aumentado em incrementos de 1TB. Pode ser preferível construir o grupo de discos a partir de oito LUNs de 500GB cada, para que o grupo de discos possa ser aumentado em incrementos menores.

A prática de estabelecer um tamanho de LUN padrão universal é desencorajada porque isso pode complicar a capacidade de gerenciamento. Por exemplo, um tamanho de LUN padrão de 100GB pode funcionar bem quando um banco de dados ou datastore está no intervalo de 1TB a 2TB, mas um banco de dados ou datastore de 20TB GB de tamanho exigiria 200 LUNs. Isso significa que os tempos de reinicialização do servidor são mais longos, há mais objetos para gerenciar nas várias UIs e produtos como SnapCenter devem executar a descoberta em muitos objetos. O uso de menos LUNs maiores evita esses problemas.

- A contagem de LUN é mais importante do que o tamanho do LUN.
- O tamanho do LUN é controlado principalmente pelos requisitos de contagem de LUN.
- Evite criar mais LUNs do que o necessário.

Contagem de LUN

Ao contrário do tamanho do LUN, a contagem de LUN afeta o desempenho. O desempenho da aplicação geralmente depende da capacidade de executar e/S paralela pela camada SCSI. Como resultado, dois LUNs oferecem melhor performance do que um único LUN. Usar um LVM como Veritas VxVM, Linux LVM2 ou Oracle ASM é o método mais simples para aumentar o paralelismo.

Os clientes da NetApp geralmente experimentaram o mínimo de benefícios ao aumentar o número de LUNs além de dezasseis, embora o teste de ambientes 100% SSD com e/S aleatória muito pesada tenha demonstrado melhorias adicionais em até 64 LUNs.

A NetApp recomenda o seguinte:

Em geral, quatro a dezesseis LUNs são suficientes para atender às necessidades de e/S de qualquer workload de banco de dados. Menos de quatro LUNs podem criar limitações de desempenho devido às limitações nas implementações de SCSI de host.

Colocação de LUN

O posicionamento ideal de LUNs de banco de dados nos volumes do ONTAP depende principalmente de como vários recursos do ONTAP serão usados.

Volumes

Um ponto comum de confusão com os clientes novos no ONTAP é o uso de FlexVols, comumente chamado de simplesmente "volumes".

Um volume não é um LUN. Esses termos são usados de forma sinônima com muitos outros produtos de fornecedores, incluindo provedores de nuvem. O ONTAP volumes é simplesmente volumes de gerenciamento. Eles não servem dados por si mesmos, nem ocupam espaço. Eles são contentores para arquivos ou LUNs e existem para melhorar e simplificar a capacidade de gerenciamento, especialmente em escala.

Volumes e LUNs

Os LUNs relacionados normalmente são colocalizados em um único volume. Por exemplo, um banco de dados que requer 10 LUNs normalmente teria todos os 10 LUNs colocados no mesmo volume.

- Usar uma proporção de 1:1:1, de LUNs para volumes, ou seja, um LUN por volume, é **não** uma prática recomendada formal.
- Em vez disso, os volumes devem ser vistos como contêineres para workloads ou conjuntos de dados. Pode haver um único LUN por volume, ou pode haver muitos. A resposta certa depende dos requisitos de gerenciabilidade.
- Espalhar LUNs em um número desnecessário de volumes pode levar a problemas adicionais de sobrecarga e agendamento para operações, como operações de snapshot, números excessivos de objetos exibidos na IU e resultar em alcançar limites de volume da plataforma antes que o limite de LUN seja atingido.

Volumes, LUNs e instantâneos

As políticas e programações de snapshot são colocadas no volume, não no LUN. Um conjunto de dados que consiste em 10 LUNs exigiria apenas uma única política de snapshot quando os LUNs estiverem colocalizados no mesmo volume.

Além disso, a localização conjunta de todos os LUNs relacionados para um determinado conjunto de dados em um único volume proporciona operações de snapshot atômico. Por exemplo, um banco de dados que residia em 10 LUNs ou um ambiente de aplicativo baseado em VMware que consiste em 10 sistemas operacionais diferentes pode ser protegido como um único objeto consistente se os LUNs subjacentes forem todos colocados em um único volume. Se forem colocados em volumes diferentes, os instantâneos podem ou não estar 100% sincronizados, mesmo que programados ao mesmo tempo.

Em alguns casos, um conjunto relacionado de LUNs pode precisar ser dividido em dois volumes diferentes devido aos requisitos de recuperação. Por exemplo, um banco de dados pode ter quatro LUNs para datafiles e dois LUNs para logs. Nesse caso, um volume de arquivo de dados com 4 LUNs e um volume de log com 2

LUNs podem ser a melhor opção. A razão é recuperabilidade independente. Por exemplo, o volume do arquivo de dados poderia ser restaurado seletivamente para um estado anterior, o que significa que todos os quatro LUNs seriam revertidos para o estado do instantâneo, enquanto o volume de log com seus dados críticos não seria afetado.

Volumes, LUNs e SnapMirror

As políticas e operações do SnapMirror são, como operações de snapshot, executadas no volume, e não no LUN.

A realocação de LUNs relacionados em um único volume permite criar uma única relação do SnapMirror e atualizar todos os dados contidos com uma única atualização. Tal como acontece com instantâneos, a atualização também será uma operação atômica. O destino do SnapMirror teria a garantia de ter uma réplica pontual das LUNs de origem. Se os LUNs foram espalhados por vários volumes, as réplicas podem ou não ser consistentes umas com as outras.

Volumes, LUNs e QoS

Embora a QoS possa ser aplicada seletivamente a LUNs individuais, geralmente é mais fácil defini-la no nível do volume. Por exemplo, todos os LUNs usados pelos convidados em um determinado servidor ESX podem ser colocados em um único volume e, em seguida, uma política de QoS adaptável do ONTAP pode ser aplicada. O resultado é um limite de IOPS por TB com dimensionamento automático que se aplica a todos os LUNs.

Da mesma forma, se um banco de dados exigisse 100K IOPS e 10 LUNs ocupados, seria mais fácil definir um único limite de 100K IOPS em um único volume do que definir 10 limites de 10K IOPS individuais, um em cada LUN.

Esquemas de vários volumes

Há alguns casos em que a distribuição de LUNs em vários volumes pode ser benéfica. O principal motivo é o striping do controlador. Por exemplo, um sistema de storage de HA pode hospedar um único banco de dados em que o potencial de processamento e armazenamento em cache completo de cada controlador é necessário. Nesse caso, um design típico seria colocar metade dos LUNs em um único volume no controlador 1 e a outra metade dos LUNs em um único volume no controlador 2.

Da mesma forma, o striping do controlador pode ser usado para balanceamento de carga. Um sistema de HA que hospedava 100 bancos de dados de 10 LUNs cada um pode ser projetado no local em que cada banco de dados recebe um volume de 5 LUN em cada uma das duas controladoras. O resultado é o carregamento simétrico garantido de cada controlador à medida que bancos de dados adicionais são provisionados.

No entanto, nenhum desses exemplos envolve uma proporção de volume para LUN de 1:1:1. O objetivo continua a ser otimizar a capacidade de gerenciamento, colocando LUNs relacionados em volumes.

Um exemplo em que uma taxa de 1:1 LUN para volume faz sentido é a Conteinerização, onde cada LUN pode realmente representar uma única carga de trabalho e precisa ser gerenciado individualmente. Nesses casos, uma proporção de 1:1 pode ser ótima.

Redimensionamento LUN e redimensionamento LVM

Quando um sistema de arquivos baseado em SAN atingiu seu limite de capacidade, há duas opções para aumentar o espaço disponível:

- Aumente o tamanho dos LUNs

- Adicione um LUN a um grupo de volumes existente e aumente o volume lógico contido

Embora o redimensionamento LUN seja uma opção para aumentar a capacidade, geralmente é melhor usar um LVM, incluindo o Oracle ASM. Uma das principais razões pelas quais LVMs existem é evitar a necessidade de um redimensionamento LUN. Com uma LVM, vários LUNs são colados em um pool virtual de storage. Os volumes lógicos esculpidos fora deste pool são gerenciados pela LVM e podem ser facilmente redimensionados. Um benefício adicional é evitar hotspots em uma determinada unidade, distribuindo um determinado volume lógico em todos os LUNs disponíveis. A migração transparente geralmente pode ser realizada usando o gerenciador de volumes para realocar as extensões subjacentes de um volume lógico para novos LUNs.

LVM striping (distribuição LVM)

A distribuição de LVM refere-se à distribuição de dados entre vários LUNs. O resultado é uma melhoria significativa na performance de muitos bancos de dados.

Antes da era das unidades flash, a distribuição foi usada para ajudar a superar as limitações de desempenho das unidades giratórias. Por exemplo, se um sistema operacional precisar executar uma operação de leitura 1MB, ler que 1MB TB de dados de uma única unidade exigiria muita busca e leitura de cabeça de unidade, pois o 1MB é transferido lentamente. Se esse 1MB TB de dados fosse distribuído em 8 LUNs, o sistema operacional poderia emitir oito operações de leitura 128K em paralelo e reduzir o tempo necessário para concluir a transferência 1MB.

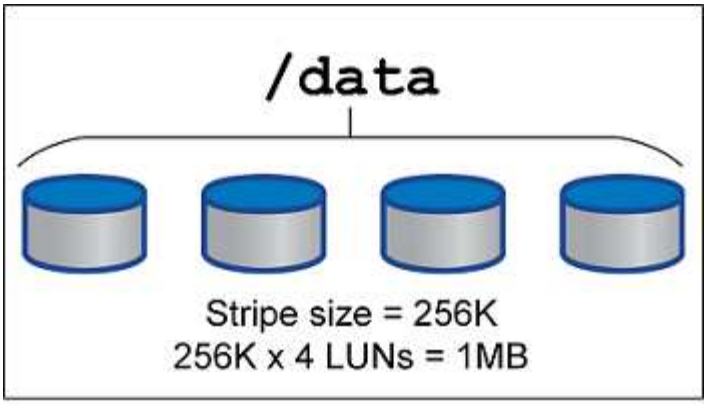
Riscar com unidades giratórias foi mais difícil porque o padrão de e/S tinha que ser conhecido com antecedência. Se o striping não foi ajustado corretamente para os padrões de e/S verdadeiros, as configurações listradas podem danificar o desempenho. Com os bancos de dados Oracle e, especialmente, com configurações all-flash, a distribuição é muito mais fácil de configurar e, comprovadamente, aumentou significativamente o desempenho.

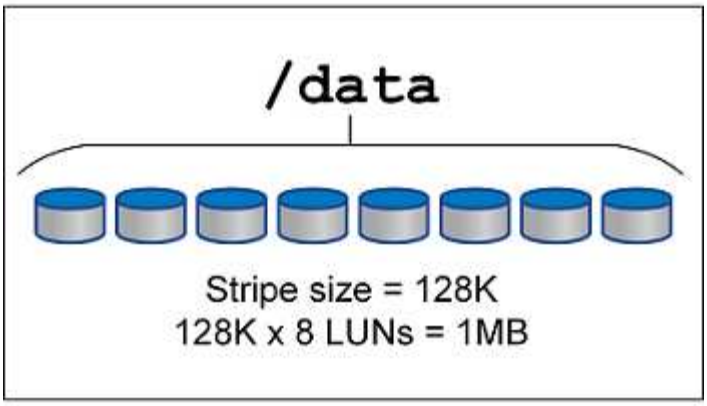
Gerenciadores de volume lógicos, como o Oracle ASM stripe por padrão, mas o LVM de SO nativo não. Alguns deles unem vários LUNs como um dispositivo concatenado, o que resulta em datafiles que existem em um e apenas um dispositivo LUN. Isso causa pontos quentes. Outras implementações de LVM são padrão para extensões distribuídas. Isso é semelhante ao striping, mas é mais grosseiro. Os LUNs no grupo de volumes são cortados em pedaços grandes, chamados de extensões e normalmente medidos em muitos megabytes, e os volumes lógicos são então distribuídos por essas extensões. O resultado é a e/S aleatória contra um arquivo deve ser bem distribuída entre LUNs, mas as operações de e/S sequenciais não são tão eficientes quanto poderiam ser.

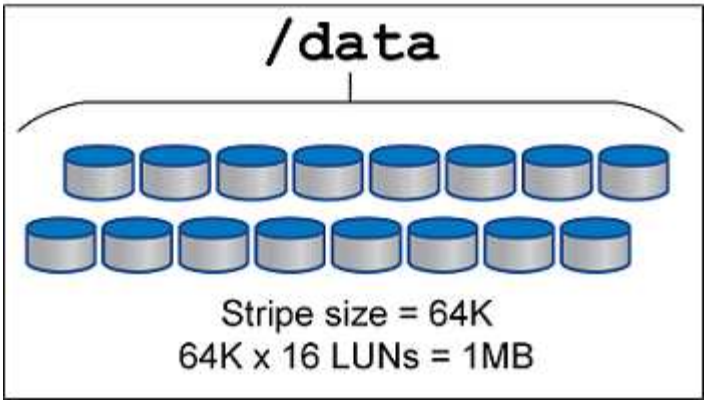
A e/S de aplicativos intensivos quase sempre é (a) em unidades do tamanho básico do bloco ou (b) um megabyte.

O principal objetivo de uma configuração de distribuição é garantir que a e/S de arquivo único possa ser executada como uma única unidade, e as e/S de vários blocos, que devem ter 1MB MB de tamanho, possam ser paralelizadas uniformemente em todos os LUNs no volume distribuído. Isso significa que o tamanho do stripe não deve ser menor que o tamanho do bloco do banco de dados e o tamanho do stripe multiplicado pelo número de LUNs deve ser 1MB.

A figura a seguir mostra três opções possíveis para o tamanho da faixa e ajuste de largura. O número de LUNs é selecionado para atender aos requisitos de performance conforme descrito acima, mas, em todos os casos, o total de dados em um único stripe é de 1MB.







NFS

Visão geral

A NetApp fornece storage NFS de nível empresarial há mais de 30 anos, e seu uso está crescendo cada vez mais, levando em consideração a simplicidade de suas infraestruturas baseadas em nuvem.

O protocolo NFS inclui várias versões com requisitos variados. Para obter uma descrição completa da configuração NFS com o ONTAP, "TR-4067 NFS nas práticas recomendadas da ONTAP"consulte . As seções a seguir abrangem alguns dos requisitos mais críticos e erros comuns do usuário.

Versões de NFS

O cliente do sistema operacional NFS deve ter suporte do NetApp.

- O NFSv3 é suportado com os sistemas operacionais que seguem o padrão NFSv3.

- O NFSv3 é compatível com o cliente Oracle DNFS.
- O NFSv4 é compatível com todos os sistemas operacionais que seguem o padrão NFSv4.
- NFSv4,1 e NFSv4,2 requerem suporte específico ao SO. Consulte o "NetApp IMT" para ver os SO suportados.
- O suporte ao Oracle DNFS para NFSv4,1 requer o Oracle 12.2.0.2 ou superior.

> "Matriz de suporte NetApp"O para NFSv3 e NFSv4 não inclui sistemas operacionais específicos. Todos os SO que obedecem ao RFC são geralmente suportados. Ao pesquisar no IMT on-line para suporte a NFSv3 ou NFSv4, não selecione um sistema operacional específico porque não haverá correspondências exibidas. Todos os SO são implicitamente suportados pela política geral.

Tabelas de slots TCP do Linux NFSv3

As tabelas de slot TCP são equivalentes a NFSv3 mm de profundidade de fila do adaptador de barramento do host (HBA). Essas tabelas controlam o número de operações NFS que podem ficar pendentes de uma só vez. O valor padrão é geralmente 16, o que é muito baixo para um desempenho ideal. O problema oposto ocorre em kernels Linux mais recentes, que podem aumentar automaticamente o limite da tabela de slots TCP para um nível que satura o servidor NFS com solicitações.

Para um desempenho ideal e para evitar problemas de desempenho, ajuste os parâmetros do kernel que controlam as tabelas de slots TCP.

Executar o `sysctl -a | grep tcp.*.slot_table` comando e respeitar os seguintes parâmetros:

```
# sysctl -a | grep tcp.*.slot_table
sunrpc.tcp_max_slot_table_entries = 128
sunrpc.tcp_slot_table_entries = 128
```

Todos os sistemas Linux devem incluir `sunrpc.tcp_slot_table_entries`, mas apenas alguns incluem `sunrpc.tcp_max_slot_table_entries`. Ambos devem ser definidos para 128.

> A falha em definir esses parâmetros pode ter efeitos significativos no desempenho. Em alguns casos, o desempenho é limitado porque o sistema operacional linux não está emitindo e/S suficiente Em outros casos, as latências de e/S aumentam à medida que o sistema operacional linux tenta emitir mais e/S do que pode ser reparado.

ADR e NFS

Alguns clientes relataram problemas de desempenho resultantes de uma quantidade excessiva de e/S nos dados `ADR` no local. O problema geralmente não ocorre até que muitos dados de desempenho tenham sido acumulados. A razão para a e/S excessiva é desconhecida, mas este problema parece ser um resultado de processos Oracle repetidamente verificando o diretório de destino para mudanças.

A remoção `noac` das opções e/ou `actimeo=0` montagem permite que o armazenamento em cache do sistema operacional do host ocorra e reduz os níveis de e/S de storage.

A NetApp recomenda não colocar `ADR` dados em um sistema de arquivos com `noac` ou `actimeo=0` porque problemas de desempenho são prováveis. Separe `ADR` os dados em um ponto de montagem diferente, se necessário.

nfs-rootonly e mount-rootonly

O ONTAP inclui uma opção NFS chamada `nfs-rootonly` que controla se o servidor aceita conexões de tráfego NFS de portas altas. Como medida de segurança, apenas o usuário raiz tem permissão para abrir conexões TCP/IP usando uma porta de origem abaixo de 1024, porque essas portas são normalmente reservadas para uso do sistema operacional, não para processos de usuário. Essa restrição ajuda a garantir que o tráfego NFS seja de um cliente NFS do sistema operacional real e não de um processo mal-intencionado emulando um cliente NFS. O cliente Oracle DNFS é um driver de espaço de usuário, mas o processo é executado como root, portanto geralmente não é necessário alterar o valor de `nfs-rootonly`. As conexões são feitas a partir de portas baixas.

A `mount-rootonly` opção só se aplica a NFSv3. Ele controla se a chamada DE MONTAGEM RPC será aceita a partir de portas maiores que 1024. Quando o DNFS é usado, o cliente está novamente executando como root, para que ele possa abrir portas abaixo de 1024. Este parâmetro não tem efeito.

Os processos de abertura de conexões com DNFS em NFS versões 4,0 e superiores não são executados como raiz e, portanto, exigem portas acima de 1024. O `nfs-rootonly` parâmetro deve ser definido como desativado para que o DNFS conclua a conexão.

Se `nfs-rootonly` o estiver ativado, o resultado será uma parada durante a fase de montagem abrindo conexões DNFS. A saída sqlplus é semelhante a esta:

```
SQL>startup
ORACLE instance started.
Total System Global Area 4294963272 bytes
Fixed Size                  8904776 bytes
Variable Size             822083584 bytes
Database Buffers         3456106496 bytes
Redo Buffers                7868416 bytes
```

O parâmetro pode ser alterado da seguinte forma:

```
Cluster01::> nfs server modify -nfs-rootonly disabled
```

Em situações raras, você pode precisar alterar tanto nfs-rootonly quanto mount-rootonly para desabilitado. Se um servidor estiver gerenciando um número extremamente grande de conexões TCP, é possível que nenhuma porta abaixo de 1024 esteja disponível e o sistema operacional seja forçado a usar portas mais altas. Esses dois parâmetros ONTAP precisariam ser alterados para permitir que a conexão fosse concluída.

Políticas de exportação de NFS: Superusuário e setuid

Se os binários Oracle estiverem localizados em um compartilhamento NFS, a política de exportação deverá incluir permissões de superusuário e setuid.

As exportações de NFS compartilhadas usadas para serviços de arquivos genéricos, como diretórios home do usuário, geralmente esmagam o usuário raiz. Isso significa que uma solicitação do usuário root em um host que montou um sistema de arquivos é remapeada como um usuário diferente com Privileges inferior. Isso ajuda a proteger os dados, impedindo que um usuário root em um servidor específico acesse dados no servidor compartilhado. O bit setuid também pode ser um risco de segurança em um ambiente compartilhado. O bit setuid permite que um processo seja executado como um usuário diferente do usuário que invoca o comando. Por exemplo, um script shell que era de propriedade do root com o bit setuid é executado como root. Se esse script shell pudesse ser alterado por outros usuários, qualquer usuário que não seja root poderá emitir um comando como root atualizando o script.

Os binários Oracle incluem arquivos de propriedade do root e usam o bit setuid. Se os binários Oracle estiverem instalados em um compartilhamento NFS, a política de exportação deverá incluir as permissões de superusuário e setuid apropriadas. No exemplo abaixo, a regra inclui tanto `allow-suid` e permite `superuser` o acesso (raiz) para clientes NFS usando autenticação de sistema.

```
Cluster01::> export-policy rule show -vserver vserver1 -policyname orabin
-fields allow-suid,superuser
vserver   policyname ruleindex superuser allow-suid
--------- ---------- --------- --------- ----------
vserver1  orabin     1         sys       true
```

Configuração NFSv4/4,1

Para a maioria das aplicações, há muito pouca diferença entre NFSv3 e NFSv4. A e/S da aplicação geralmente é muito simples e/S e não se beneficia significativamente de alguns dos recursos avançados disponíveis no NFSv4. Versões mais altas do NFS não devem ser vistas como uma "atualização" da perspectiva do storage de banco de dados, mas sim como versões do NFS que incluem recursos adicionais. Por exemplo, se a segurança de ponta a ponta do modo de privacidade Kerberos (krb5p) for necessária, então NFSv4 será necessário.

A NetApp recomenda usar o NFSv4,1 se forem necessários recursos do NFSv4. Há algumas melhorias funcionais no protocolo NFSv4 em NFSv4,1 que melhoram a resiliência em certos casos de borda.

Mudar para NFSv4 é mais complicado do que simplesmente mudar as opções de montagem de vers-3 para vers-4,1. Uma explicação mais completa da configuração do NFSv4 com o ONTAP, incluindo orientações sobre a configuração do sistema operacional, "TR-4067 NFS nas práticas recomendadas da ONTAP"consulte . As secções seguintes deste TR explicam alguns dos requisitos básicos para a utilização do NFSv4.

Domínio NFSv4

Uma explicação completa da configuração NFSv4/4,1 está além do escopo deste documento, mas um problema comumente encontrado é uma incompatibilidade no mapeamento de domínio. De um ponto de vista sysadmin, os sistemas de arquivos NFS parecem se comportar normalmente, mas os aplicativos relatam erros sobre permissões e/ou setuid em determinados arquivos. Em alguns casos, os administradores concluíram incorretamente que as permissões dos binários do aplicativo foram danificadas e executaram comandos chown ou chmod quando o problema real era o nome do domínio.

O nome de domínio NFSv4 é definido no ONTAP SVM:

```
Cluster01::> nfs server show -fields v4-id-domain
vserver   v4-id-domain
--------- ------------
vserver1  my.lab
```

O nome de domínio NFSv4 no host é definido em `/etc/idmap.cfg`

```
[root@host1 etc]# head /etc/idmapd.conf
[General]
#Verbosity = 0
# The following should be set to the local NFSv4 domain name
# The default is the host's DNS domain name.
Domain = my.lab
```

Os nomes de domínio devem corresponder. Se não o fizerem, erros de mapeamento semelhantes aos seguintes aparecem em `/var/log/messages`:

```
Apr 12 11:43:08 host1 nfsidmap[16298]: nss_getpwnam: name 'root@my.lab'
does not map into domain 'default.com'
```

Binários de aplicativos, como binários de banco de dados Oracle, incluem arquivos de propriedade do root com o bit setuid, o que significa que uma incompatibilidade nos nomes de domínio NFSv4 causa falhas na inicialização do Oracle e um aviso sobre a propriedade ou permissões de um arquivo chamado `oradism`, que está localizado no `$ORACLE_HOME/bin` diretório. Deve aparecer da seguinte forma:

```
[root@host1 etc]# ls -l /orabin/product/19.3.0.0/dbhome_1/bin/oradism
-rwsr-x--- 1 root oinstall 147848 Apr 17  2019
/orabin/product/19.3.0.0/dbhome_1/bin/oradism
```

Se este arquivo aparecer com a propriedade de ninguém, pode haver um problema de mapeamento de domínio NFSv4.

```
[root@host1 bin]# ls -l oradism
-rwsr-x--- 1 nobody oinstall 147848 Apr 17  2019 oradism
```

Para corrigir isso, verifique o `/etc/idmap.cfg` arquivo na configuração v4-id-domain no ONTAP e certifique-se de que eles sejam consistentes. Se não estiverem, faça as alterações necessárias, execute `nfsidmap -c` e aguarde um momento para que as alterações se propaguem. A propriedade do arquivo deve então ser devidamente reconhecida como raiz. Se um usuário tivesse tentado executar `chown root` esse arquivo antes que a configuração dos domínios NFS fosse corrigida, talvez seja necessário executar `chown root` novamente.

Oracle Direct NFS (DNFS)

Os bancos de dados Oracle podem usar o NFS de duas maneiras.

Primeiro, ele pode usar um sistema de arquivos montado usando o cliente NFS nativo que faz parte do sistema operacional. Isso às vezes é chamado de kernel NFS, ou kNFS. O sistema de arquivos NFS é montado e usado pelo banco de dados Oracle exatamente o mesmo que qualquer outro aplicativo usaria um sistema de arquivos NFS.

O segundo método é o Oracle Direct NFS (DNFS). Esta é uma implementação do padrão NFS no software de banco de dados Oracle. Ele não altera a maneira como os bancos de dados Oracle são configurados ou gerenciados pelo DBA. Desde que o próprio sistema de armazenamento tenha as configurações corretas, o uso do DNFS deve ser transparente para o grupo DBA e para os usuários finais.

Um banco de dados com o recurso DNFS habilitado ainda tem os sistemas de arquivos NFS usuais montados. Uma vez que o banco de dados está aberto, o banco de dados Oracle abre um conjunto de sessões TCP/IP e executa operações NFS diretamente.

NFS direto

O principal valor do NFS direto da Oracle é ignorar o cliente NFS do host e executar operações de arquivos NFS diretamente em um servidor NFS. A ativação da TI requer apenas a alteração da biblioteca do Oracle Disk Manager (ODM). As instruções para este processo são fornecidas na documentação da Oracle.

O uso do DNFS resulta em uma melhoria significativa no desempenho de e/S e diminui a carga no host e no sistema de armazenamento, pois a e/S é realizada da maneira mais eficiente possível.

Além disso, o Oracle DNFS inclui uma **opção** para multipathing de interface de rede e tolerância a falhas. Por exemplo, duas interfaces 10Gb podem ser Unidas para oferecer 20Gb Gbps de largura de banda. Uma falha de uma interface faz com que a I/o seja tentada novamente na outra interface. A operação geral é muito semelhante ao multipathing FC. Multipathing era comum anos atrás, quando 1GB ethernet era o padrão mais comum. Uma NIC 10Gb é suficiente para a maioria das cargas de trabalho Oracle, mas se for necessário mais, 10Gb NICs podem ser colados.

Quando o DNFS é usado, é fundamental que todos os patches descritos no Oracle Doc 1495104,1 sejam instalados. Se um patch não puder ser instalado, o ambiente deve ser avaliado para garantir que os bugs descritos nesse documento não causem problemas. Em alguns casos, a incapacidade de instalar os patches necessários impede o uso do DNFS.

Não use DNFS com qualquer tipo de resolução de nome de round-robin, incluindo DNS, DDNS, NIS ou qualquer outro método. Isso inclui o recurso de balanceamento de carga DNS disponível no ONTAP. Quando um banco de dados Oracle usando DNFS resolve um nome de host para um endereço IP, ele não deve ser alterado em pesquisas subsequentes. Isso pode resultar em falhas de banco de dados Oracle e possível corrupção de dados.

Ativar DNFS

O Oracle DNFS pode trabalhar com o NFSv3 sem necessidade de configuração além de ativar a biblioteca DNFS (consulte a documentação Oracle para o comando específico necessário), mas se o DNFS não conseguir estabelecer conetividade, ele pode reverter silenciosamente para o cliente NFS do kernel. Se isso acontecer, o desempenho pode ser gravemente afetado.

Se você deseja usar a multiplexação DNFS em várias interfaces, com NFSv4.X, ou usar criptografia, você deve configurar um arquivo oranfstab. A sintaxe é extremamente rigorosa. Pequenos erros no arquivo podem resultar em suspensão de inicialização ou ignorar o arquivo oranfstab.

No momento da escrita, o multipathing DNFS não funciona com o NFSv4,1 com versões recentes do Oracle Database. Um arquivo oranfstab que especifica NFSv4,1 como um protocolo só pode usar uma instrução de caminho único para uma determinada exportação. O motivo é que o ONTAP não oferece suporte ao entroncamento ClientID. Patches do banco de dados Oracle para resolver essa limitação podem estar disponíveis no futuro.

A única maneira de ter certeza de que o DNFS está operando como esperado é consultar as tabelas dnfs.

Abaixo está um exemplo de arquivo oranfstab localizado em /etc. Este é um dos vários locais que um arquivo oranfstab pode ser colocado.

```
[root@jfs11 trace]# cat /etc/oranfstab
server: NFSv3test
path: jfs_svmdr-nfs1
path: jfs_svmdr-nfs2
export: /dbf mount: /oradata
export: /logs mount: /logs
nfs_version: NFSv3
```

O primeiro passo é verificar se o DNFS está operacional para os sistemas de arquivos especificados:

```
SQL> select dirname,nfsversion from v$dnfs_servers;

DIRNAME
-----------------------------------
NFSVERSION
----------------
/logs
NFSv3.0

/dbf
NFSv3.0
```

Essa saída indica que o DNFS está em uso com esses dois sistemas de arquivos, mas ele não significa que o oranfstab esteja operacional. Se um erro estivesse presente, o DNFS teria descoberto automaticamente os sistemas de arquivos NFS do host e você ainda poderá ver a mesma saída deste comando.

Multipathing pode ser verificado da seguinte forma:

```
SQL> select svrname,path,ch_id from v$dnfs_channels;

SVRNAME
-----------------------------------
PATH
-----------------------------------
     CH_ID
```

```
----------
NFSv3test
jfs_svmdr-nfs1
         0

NFSv3test
jfs_svmdr-nfs2
         1

SVRNAME
-------------------------------------
PATH
-------------------------------------
     CH_ID
----------

NFSv3test
jfs_svmdr-nfs1
         0

NFSv3test
jfs_svmdr-nfs2

[output truncated]

SVRNAME
-------------------------------------
PATH
-------------------------------------
     CH_ID
----------
NFSv3test
jfs_svmdr-nfs2
         1

NFSv3test
jfs_svmdr-nfs1
         0

SVRNAME
-------------------------------------
PATH
-------------------------------------
     CH_ID
----------
```

```
NFSv3test
jfs_svmdr-nfs2
         1


66 rows selected.
```

Estas são as conexões que o DNFS está usando. Dois caminhos e canais são visíveis para cada entrada SVRNAME. Isso significa que o multipathing está funcionando, o que significa que o arquivo oranfstab foi reconhecido e processado.

Acesso direto ao NFS e ao sistema de arquivos do host

O uso do DNFS pode ocasionalmente causar problemas para aplicativos ou atividades do usuário que dependem dos sistemas de arquivos visíveis montados no host porque o cliente DNFS acessa o sistema de arquivos fora da banda do sistema operacional do host. O cliente DNFS pode criar, excluir e modificar arquivos sem o conhecimento do sistema operacional.

Quando as opções de montagem para bancos de dados de instância única são usadas, elas permitem o armazenamento em cache de atributos de arquivo e diretório, o que também significa que o conteúdo de um diretório é armazenado em cache. Portanto, o DNFS pode criar um arquivo, e há um curto atraso antes que o sistema operacional releia o conteúdo do diretório e o arquivo se torne visível para o usuário. Isso geralmente não é um problema, mas, em raras ocasiões, utilitários como SAP BR*Tools podem ter problemas. Se isso acontecer, solucione o problema alterando as opções de montagem para usar as recomendações do Oracle RAC. Essa alteração resulta na desativação de todo o cache do host.

Altere apenas as opções de montagem quando (a) DNFS for usado e (b) um problema resulta de um atraso na visibilidade do arquivo. Se o DNFS não estiver em uso, o uso das opções de montagem do Oracle RAC em um banco de dados de instância única resulta em desempenho degradado.

Veja a nota sobre `nosharecache` o in "Opções de montagem em NFS do Linux" para um problema DNFS específico do Linux que pode produzir resultados incomuns.

Locações e bloqueios de NFS

O NFSv3 está sem monitoração de estado. Isso efetivamente significa que o servidor NFS (ONTAP) não controla quais sistemas de arquivos são montados, por quem, ou quais bloqueios estão realmente no lugar.

O ONTAP tem alguns recursos que gravarão tentativas de montagem para que você tenha uma ideia de quais clientes podem estar acessando dados, e pode haver bloqueios de consultoria presentes, mas essa informação não é garantida para ser 100% completa. Ele não pode ser concluído, porque o rastreamento do estado do cliente NFS não faz parte do padrão NFSv3.

NFSv4 declaração

Em contraste, NFSv4 é stateful. O servidor NFSv4 rastreia quais clientes estão usando quais sistemas de arquivos, quais arquivos existem, quais arquivos e/ou regiões de arquivos estão bloqueados, etc. isso significa que precisa haver comunicação regular entre um servidor NFSv4 para manter os dados de estado atuais.

Os estados mais importantes que estão sendo gerenciados pelo servidor NFS são NFSv4 bloqueios e NFSv4 arrendamentos, e eles estão muito interligados. Você precisa entender como cada um funciona por si mesmo,

e como eles se relacionam uns com os outros.

NFSv4 fechaduras

Com o NFSv3, os bloqueios são consultivos. Um cliente NFS ainda pode modificar ou excluir um arquivo "bloqueado". Um bloqueio NFSv3 não expira por si só, ele deve ser removido. Isso cria problemas. Por exemplo, se você tiver um aplicativo em cluster que crie NFSv3 bloqueios e um dos nós falhar, o que você faz? Você pode codificar o aplicativo nos nós sobreviventes para remover os bloqueios, mas como você sabe que isso é seguro? Talvez o nó "falhou" esteja operacional, mas não esteja se comunicando com o restante do cluster?

Com NFSv4, os bloqueios têm uma duração limitada. Desde que o cliente que detém os bloqueios continue a efetuar o check-in com o servidor NFSv4, nenhum outro cliente tem permissão para adquirir esses bloqueios. Se um cliente não conseguir efetuar o check-in com o NFSv4, os bloqueios eventualmente serão revogados pelo servidor e outros clientes poderão solicitar e obter bloqueios.

NFSv4 arrendamentos

NFSv4 bloqueios estão associados a um leasing de NFSv4. Quando um cliente NFSv4 estabelece uma conexão com um servidor NFSv4, ele recebe um leasing. Se o cliente obtém um bloqueio (existem muitos tipos de bloqueios), então o bloqueio está associado ao leasing.

Este leasing tem um tempo limite definido. Por padrão, o ONTAP definirá o valor de tempo limite para 30 segundos:

```
Cluster01::*> nfs server show -vserver vserver1 -fields v4-lease-seconds

vserver   v4-lease-seconds
--------- ----------------
vserver1  30
```

Isso significa que um cliente NFSv4 precisa fazer o check-in com o servidor NFSv4 a cada 30 segundos para renovar suas locações.

A locação é renovada automaticamente por qualquer atividade, portanto, se o cliente estiver fazendo trabalho, não há necessidade de realizar operações de adição. Se um aplicativo ficar quieto e não estiver fazendo trabalho real, ele precisará executar uma espécie de operação keep-alive (chamada de SEQUÊNCIA). É essencialmente apenas dizer "Eu ainda estou aqui, por favor, atualize meus arrendamentos."

```
 *Question:* What happens if you lose network connectivity for 31 seconds?
O NFSv3 está sem monitoração de estado. Não está à espera de comunicação
dos clientes. O NFSv4 tem estado monitorado e, uma vez decorrido esse
período de locação, o leasing expira e os bloqueios são revogados e os
arquivos bloqueados são disponibilizados para outros clientes.
```

Com o NFSv3, você pode mover cabos de rede, reinicializar switches de rede, fazer alterações de configuração e ter certeza de que nada de ruim aconteceria. Os aplicativos normalmente apenas esperariam pacientemente para que a conexão de rede funcione novamente.

Com NFSv4, você tem 30 segundos (a menos que você tenha aumentado o valor desse parâmetro dentro do

ONTAP) para concluir seu trabalho. Se você exceder isso, suas locações expiram. Normalmente, isso resulta em falhas no aplicativo.

Como exemplo, se você tiver um banco de dados Oracle e tiver uma perda de conetividade de rede (às vezes chamada de "partição de rede") que exceda o tempo limite de concessão, você irá travar o banco de dados.

Aqui está um exemplo do que acontece no log de alerta Oracle se isso acontecer:

```
2022-10-11T15:52:55.206231-04:00
Errors in file /orabin/diag/rdbms/ntap/NTAP/trace/NTAP_ckpt_25444.trc:
ORA-00202: control file: '/redo0/NTAP/ctrl/control01.ctl'
ORA-27072: File I/O error
Linux-x86_64 Error: 5: Input/output error
Additional information: 4
Additional information: 1
Additional information: 4294967295
2022-10-11T15:52:59.842508-04:00
Errors in file /orabin/diag/rdbms/ntap/NTAP/trace/NTAP_ckpt_25444.trc:
ORA-00206: error in writing (block 3, # blocks 1) of control file
ORA-00202: control file: '/redo1/NTAP/ctrl/control02.ctl'
ORA-27061: waiting for async I/Os failed
```

Se você olhar para os syslogs, você deve ver vários desses erros:

```
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
Oct 11 15:52:55 host1 kernel: NFS: nfs4_reclaim_open_state: Lock reclaim
failed!
```

As mensagens de log são geralmente o primeiro sinal de um problema, além do congelamento do aplicativo. Normalmente, você não vê nada durante a interrupção da rede porque os processos e o próprio sistema operacional estão bloqueados tentando acessar o sistema de arquivos NFS.

Os erros aparecem depois que a rede estiver operacional novamente. No exemplo acima, uma vez que a conetividade foi restabelecida, o sistema operacional tentou readquirir os bloqueios, mas era tarde demais. O arrendamento expirou e os bloqueios foram removidos. Isso resulta em um erro que se propaga até a camada Oracle e causa a mensagem no log de alerta. Você pode ver variações desses padrões dependendo da versão e configuração do banco de dados.

Em resumo, o NFSv3 tolera a interrupção da rede, mas o NFSv4 é mais sensível e impõe um período de locação definido.

E se um tempo limite de 30 segundos não for aceitável? E se você gerenciar uma rede que muda dinamicamente onde os switches são reinicializados ou os cabos são realocados e o resultado é a interrupção ocasional da rede? Você pode optar por estender o período de locação, mas se você quiser fazer isso requer uma explicação de NFSv4 períodos de carência.

NFSv4 períodos de carência

Se um servidor NFSv3 for reinicializado, ele estará pronto para servir o IO quase instantaneamente. Não estava mantendo qualquer tipo de estado sobre os clientes. O resultado é que uma operação de aquisição da ONTAP geralmente parece estar próxima de instantânea. No momento em que um controlador estiver pronto para começar a fornecer dados, ele enviará um ARP para a rede que sinaliza a alteração na topologia. Os clientes normalmente detetam isso quase instantaneamente e os dados continuam fluindo.

NFSv4, no entanto, irá produzir uma breve pausa. É apenas parte de como o NFSv4 funciona.

> As seções a seguir são atuais a partir do ONTAP 9.15,1, mas o comportamento de leasing e bloqueio, bem como as opções de ajuste podem mudar de versão para versão. Se você precisar ajustar os tempos de leasing/bloqueio NFSv4, consulte o suporte da NetApp para obter as informações mais recentes.

Os servidores NFSv4 precisam rastrear os arrendamentos, bloqueios e quem está usando quais dados. Se um servidor NFS picar e reiniciar, ou perder energia por um momento, ou for reiniciado durante a atividade de manutenção, o resultado será o leasing/bloqueio e outras informações do cliente serão perdidas. O servidor precisa descobrir qual cliente está usando quais dados antes de retomar as operações. É aqui que entra o período de carência.

Se você de repente ligar o seu servidor NFSv4. Quando ele voltar, os clientes que tentam retomar o IO receberão uma resposta que diz essencialmente: "Perdi informações de leasing/bloqueio. Pretende registar novamente os seus bloqueios?" Esse é o início do período de carência. O padrão é 45 segundos no ONTAP:

```
Cluster01::> nfs server show -vserver vserver1 -fields v4-grace-seconds

vserver   v4-grace-seconds
--------- ----------------
vserver1  45
```

O resultado é que, após uma reinicialização, um controlador irá pausar o IO enquanto todos os clientes recuperam seus arrendamentos e bloqueios. Quando o período de carência terminar, o servidor retomará as operações de e/S.

Esse período de carência controla a recuperação de leasing durante alterações na interface de rede, mas há um segundo período de carência que controla a recuperação durante o failover de storage, `locking.grace_lease_seconds`. Esta é uma opção de nível de nó.

```
cluster01::> node run [node names or *] options
locking.grace_lease_seconds
```

Por exemplo, se você precisasse frequentemente executar failovers de LIF e precisasse reduzir o período de carência, você mudaria `v4-grace-seconds`. Se você quisesse melhorar o tempo de retomada de e/S durante o failover da controladora, seria necessário alterar `locking.grace_lease_seconds`o .

Apenas altere estes valores com cautela e depois de compreender completamente os riscos e consequências. As pausas de e/S envolvidas com operações de failover e migração com o NFSv4.X não podem ser totalmente evitadas. Os períodos de bloqueio, leasing e carência fazem parte da RFC do NFS. Para muitos clientes, o NFSv3 é preferível porque os tempos de failover são mais rápidos.

Prazos de concessão vs períodos de carência

O período de carência e o período de leasing estão ligados. Como mencionado acima, o tempo limite de leasing padrão é de 30 segundos, o que significa que NFSv4 clientes devem fazer check-in com o servidor pelo menos a cada 30 segundos ou perder seus arrendamentos e, por sua vez, seus bloqueios. O período de carência existe para permitir que um servidor NFS reconstrua dados de concessão/bloqueio e o padrão é de 45 segundos. O período de carência deve ser superior ao período de locação. Isso garante que um ambiente cliente NFS projetado para renovar contratos de arrendamento pelo menos a cada 30 segundos terá a capacidade de fazer check-in com o servidor após uma reinicialização. Um período de carência de 45 segundos garante que todos os clientes que esperam renovar seus arrendamentos pelo menos a cada 30 segundos definitivamente tenham a oportunidade de fazê-lo.

Se um tempo limite de 30 segundos não for aceitável, você pode optar por estender o período de locação.

Se você quiser aumentar o tempo limite de leasing para 60 segundos para suportar uma interrupção de rede de 60 segundos, você também terá que aumentar o período de carência. Isso significa que você terá pausas de e/S mais longas durante o failover de controladora.

Isso normalmente não deve ser um problema. Os usuários típicos atualizam somente as controladoras ONTAP uma ou duas vezes por ano, e o failover não planejado devido a falhas de hardware é extremamente raro. Além disso, se você tivesse uma rede em que uma interrupção de rede de 60 segundos era uma possibilidade preocupante, e você precisasse do tempo limite da concessão para 60 segundos, provavelmente você não obteria o failover raro do sistema de storage, resultando em uma pausa de 61 segundos também. Você já reconheceu que tem uma rede que está pausando por mais de 60 segundos com bastante frequência.

Armazenamento em cache NFS

A presença de qualquer uma das seguintes opções de montagem faz com que o cache do host seja desativado:

```
cio, actimeo=0, noac, forcedirectio
```

Essas configurações podem ter um efeito negativo grave na velocidade de instalação do software, patches e operações de backup/restauração. Em alguns casos, especialmente com aplicações em cluster, essas opções são necessárias como resultado inevitável da necessidade de fornecer coerência de cache em todos os nós do cluster. Em outros casos, os clientes usam erroneamente esses parâmetros e o resultado é danos desnecessários no desempenho.

Muitos clientes removem temporariamente essas opções de montagem durante a instalação ou o patch dos binários da aplicação. Essa remoção pode ser realizada com segurança se o usuário verificar que nenhum outro processo está usando ativamente o diretório de destino durante o processo de instalação ou patch.

Tamanhos de transferência NFS

Por padrão, o ONTAP limita os tamanhos de e/S de NFS a 64K.

A e/S aleatória com a maioria dos aplicativos e bancos de dados usa um tamanho de bloco muito menor que está bem abaixo do máximo 64K. A e/S de bloco grande geralmente é paralelizada, portanto o máximo de 64K GB também não é uma limitação para obter largura de banda máxima.

Existem algumas cargas de trabalho em que o máximo 64K cria uma limitação. Em particular, operações de um único processo, como operação de backup ou recuperação ou uma verificação de tabela completa de banco de dados, são executadas de forma mais rápida e eficiente se o banco de dados puder executar

menos, mas maiores I/os. O tamanho ideal de manuseio de e/S para ONTAP é 256K.

O tamanho máximo de transferência para um determinado SVM do ONTAP pode ser alterado da seguinte forma:

```
Cluster01::> set advanced
Warning: These advanced commands are potentially dangerous; use them only
when directed to do so by NetApp personnel.
Do you want to continue? {y|n}: y
Cluster01::*> nfs server modify -vserver vserver1 -tcp-max-xfer-size
262144
Cluster01::*>
```

> Nunca diminua o tamanho máximo de transferência permitido no ONTAP abaixo do valor de rsize/wsize dos sistemas de arquivos NFS atualmente instalados. Isso pode criar pendências ou até mesmo corrupção de dados com alguns sistemas operacionais. Por exemplo, se os clientes NFS estiverem atualmente definidos em um rsize/wsize de 65536, o tamanho máximo de transferência do ONTAP poderá ser ajustado entre 65536 e 1048576 sem efeito porque os próprios clientes são limitados. Reduzir o tamanho máximo de transferência abaixo de 65536 pode danificar a disponibilidade ou os dados.

NVFAIL

O NVFAIL é um recurso do ONTAP que garante a integridade durante cenários de failover catastróficos.

Os bancos de dados estão vulneráveis a corrupção durante eventos de failover de storage porque mantêm grandes caches internos. Se um evento catastrófico exigir forçar um failover de ONTAP ou forçar o switchover de MetroCluster, independentemente da integridade da configuração geral, o resultado será confirmado anteriormente que as alterações podem ser efetivamente descartadas. O conteúdo da matriz de armazenamento salta para trás no tempo, e o estado do cache do banco de dados não reflete mais o estado dos dados no disco. Essa inconsistência resulta em corrupção de dados.

O cache pode ocorrer na camada de aplicativo ou servidor. Por exemplo, uma configuração do Oracle Real Application Cluster (RAC) com servidores ativos em um site primário e remoto armazena dados em cache no Oracle SGA. Uma operação de comutação forçada que resultou em dados perdidos colocaria o banco de dados em risco de corrupção porque os blocos armazenados no SGA podem não corresponder aos blocos no disco.

Um uso menos óbvio do cache está na camada do sistema de arquivos do sistema operacional. Blocos de um sistema de arquivos NFS montado podem ser armazenados em cache no sistema operacional. Como alternativa, um sistema de arquivos em cluster baseado em LUNs localizados no site principal pode ser montado em servidores no local remoto e, mais uma vez, os dados podem ser armazenados em cache. Uma falha do NVRAM ou de uma tomada forçada ou comutação forçada nessas situações pode resultar em corrupção do sistema de arquivos.

O ONTAP protege bancos de dados e sistemas operacionais desse cenário com o NVFAIL e suas configurações associadas.

Utilitário de reclinação ASMRU (ASMRU)

O ONTAP remove com eficiência blocos zerados gravados em um arquivo ou LUN quando a compactação in-line está ativada. Utilitários como o Oracle ASM Reclamation Utility (ASRU) funcionam escrevendo zeros em extensões ASM não utilizadas.

Isso permite que os DBAs recuperem espaço na matriz de armazenamento após os dados serem excluídos. O ONTAP interceta os zeros e desaloca o espaço do LUN. O processo de recuperação é extremamente rápido porque nenhum dado está sendo gravado no sistema de storage.

Do ponto de vista do banco de dados, o grupo de discos ASM contém zeros, e a leitura dessas regiões dos LUNs resultaria em um fluxo de zeros, mas o ONTAP não armazena os zeros nas unidades. Em vez disso, são feitas mudanças simples de metadados que marcam internamente as regiões zeroadas do LUN como vazias de qualquer dado.

Por motivos semelhantes, os testes de desempenho envolvendo dados isolados não são válidos, pois blocos de zeros não são realmente processados como gravações no storage array.

Ao usar o ASRU, certifique-se de que todos os patches recomendados pela Oracle estejam instalados.

Configuração de storage em sistemas ASA R2

FC SAN

Alinhamento LUN

Alinhamento LUN refere-se a otimizar e/S em relação ao layout do sistema de arquivos subjacente.

Os sistemas ASA r2 utilizam a mesma arquitetura ONTAP que o AFF/ FAS , mas com um modelo de configuração simplificado. Os sistemas ASA r2 usam Zonas de Disponibilidade de Armazenamento (SAZ) em vez de agregados, mas os princípios de alinhamento permanecem os mesmos porque o ONTAP gerencia o layout de blocos de forma consistente em todas as plataformas. No entanto, observe estes pontos específicos da ASA:

- Os sistemas ASA r2 fornecem caminhos simétricos ativo-ativo para todos os LUNs, o que elimina preocupações com assimetria de caminho durante o alinhamento.
- As unidades de armazenamento (LUNs) são provisionadas dinamicamente por padrão; o alinhamento não altera esse comportamento.
- A reserva de snapshots e a exclusão automática de snapshots podem ser configuradas durante a criação do LUN (ONTAP 9.18.1 e posterior).

Em um sistema ONTAP, o storage é organizado em 4KB unidades. Um bloco 8KB do banco de dados ou do sistema de arquivos deve ser mapeado para exatamente dois blocos 4KB. Se um erro na configuração de LUN mudar o alinhamento em 1KB em qualquer direção, cada bloco 8KB existiria em três blocos de armazenamento 4KB diferentes em vez de dois. Esse arranjo causaria maior latência e causaria a realização de e/S adicionais no sistema de storage.

O alinhamento também afeta arquiteturas LVM. Se um volume físico dentro de um grupo de volumes lógicos for definido em todo o dispositivo da unidade (nenhuma partição é criada), o primeiro bloco 4KB no LUN se

alinha com o primeiro bloco 4KB no sistema de armazenamento. Este é um alinhamento correto. Problemas surgem com partições porque eles mudam o local inicial onde o sistema operacional usa o LUN. Desde que o deslocamento seja deslocado em unidades inteiras de 4KB, o LUN é alinhado.

Em ambientes Linux, crie grupos de volume lógicos em todo o dispositivo de unidade. Quando uma partição for necessária, verifique o alinhamento executando `fdisk -u` e verificando se o início de cada partição é um múltiplo de oito. Isso significa que a partição começa em um múltiplo de oito setores de 512 bytes, que é 4KB.

Consulte também a discussão sobre o alinhamento do bloco de compressão na "Eficiência"seção . Qualquer layout que esteja alinhado com os limites do bloco de compressão 8KBD também está alinhado com os limites 4KBD.

Avisos de desalinhamento

O log de refazer/transações do banco de dados normalmente gera e/S desalinhadas que podem causar avisos enganosos sobre LUNs desalinhados no ONTAP.

O log executa uma gravação sequencial do arquivo de log com gravações de tamanho variável. Uma operação de gravação de log que não esteja alinhada aos limites do 4KB normalmente não causa problemas de desempenho porque a próxima operação de gravação de log completa o bloco. O resultado é que o ONTAP é capaz de processar quase todas as gravações como blocos 4KB completos, mesmo que os dados em cerca de 4KB blocos tenham sido gravados em duas operações separadas.

Verifique o alinhamento usando utilitários como `sio` ou `dd` que pode gerar E/S em um tamanho de bloco definido. As estatísticas de alinhamento de E/S no sistema de armazenamento podem ser visualizadas com o `stats` comando. Ver "Verificação do alinhamento do WAFL" Para obter mais informações.

O alinhamento em ambientes Solaris é mais complicado. "Configuração do host SAN ONTAP"Consulte para obter mais informações.

> Nos ambientes Solaris x86, tenha cuidado adicional com o alinhamento adequado, pois a maioria das configurações tem várias camadas de partições. Os cortes de partição do Solaris x86 geralmente existem em cima de uma tabela de partição de Registro de inicialização principal padrão.

Boas práticas adicionais:

- Verifique as configurações de firmware e sistema operacional do HBA em relação à ferramenta NetApp Interoperability Matrix Tool (IMT).
- Utilize os utilitários do sanlun para confirmar a integridade e o alinhamento do caminho.
- Para Oracle ASM e LVM, certifique-se de que os arquivos de configuração (/etc/lvm/lvm.conf, /etc/sysconfig/oracleasm) estejam configurados corretamente para evitar problemas de alinhamento.

Dimensionamento de LUN e contagem de LUN

Selecionar o tamanho ideal de LUN e o número de LUNs a serem usados é essencial para obter o desempenho e a capacidade de gerenciamento ideais dos bancos de dados Oracle.

Um LUN é um objeto virtualizado no ONTAP que existe em todas as unidades da Zona de Disponibilidade de Armazenamento (SAZ) que hospeda os sistemas ASA r2. Consequentemente, o desempenho do LUN não é afetado pelo seu tamanho, pois o LUN aproveita todo o potencial de desempenho do SAZ, independentemente do tamanho escolhido.

Por uma questão de conveniência, os clientes podem querer usar um LUN de um tamanho específico. Por exemplo, se um banco de dados for construído em um grupo de discos LVM ou Oracle ASM composto por dois LUNs de 1TB cada, esse grupo de discos deve ser aumentado em incrementos de 1TB. Pode ser preferível construir o grupo de discos a partir de oito LUNs de 500GB cada, para que o grupo de discos possa ser aumentado em incrementos menores.

A prática de estabelecer um tamanho de LUN padrão universal é desencorajada porque isso pode complicar a capacidade de gerenciamento. Por exemplo, um tamanho de LUN padrão de 100GB pode funcionar bem quando um banco de dados ou datastore está no intervalo de 1TB a 2TB, mas um banco de dados ou datastore de 20TB GB de tamanho exigiria 200 LUNs. Isso significa que os tempos de reinicialização do servidor são mais longos, há mais objetos para gerenciar nas várias UIs e produtos como SnapCenter devem executar a descoberta em muitos objetos. O uso de menos LUNs maiores evita esses problemas.

- Considerações sobre a ASA r2:*
- O tamanho máximo de LUN para o ASA r2 é de 128 TB, o que permite um número menor de LUNs, porém maiores, sem impacto no desempenho.
- O ASA r2 usa Zonas de Disponibilidade de Armazenamento (SAZ) em vez de agregados, mas isso não altera a lógica de dimensionamento de LUN para cargas de trabalho do Oracle.
- O provisionamento dinâmico está ativado por padrão; o redimensionamento de LUNs não causa interrupções e não exige que elas sejam retiradas da internet.

Contagem de LUN

Ao contrário do tamanho do LUN, a contagem de LUN afeta o desempenho. O desempenho da aplicação geralmente depende da capacidade de executar e/S paralela pela camada SCSI. Como resultado, dois LUNs oferecem melhor performance do que um único LUN. Usar um LVM como Veritas VxVM, Linux LVM2 ou Oracle ASM é o método mais simples para aumentar o paralelismo.

Com o ASA r2, os princípios para a contagem de LUNs permanecem os mesmos que no AFF/ FAS , porque o ONTAP lida com E/S paralela de forma semelhante em todas as plataformas. No entanto, a arquitetura somente SAN do ASA r2 e os caminhos simétricos ativo-ativo garantem um desempenho consistente em todos os LUNs.

Os clientes da NetApp geralmente experimentaram o mínimo de benefícios ao aumentar o número de LUNs além de dezasseis, embora o teste de ambientes 100% SSD com e/S aleatória muito pesada tenha demonstrado melhorias adicionais em até 64 LUNs.

A NetApp recomenda o seguinte:

Em geral, de quatro a dezesseis LUNs são suficientes para atender às necessidades de E/S de qualquer carga de trabalho de banco de dados Oracle. Menos de quatro LUNs podem criar limitações de desempenho devido a limitações nas implementações SCSI do host. Aumentar o número de LUNs para além de dezesseis raramente melhora o desempenho, exceto em casos extremos (como cargas de trabalho de E/S aleatórias muito altas em SSDs).

Colocação de LUN

O posicionamento ideal dos LUNs de banco de dados em sistemas ASA r2 depende principalmente de como os diversos recursos do ONTAP serão utilizados.

Nos sistemas ASA r2, as unidades de armazenamento (LUNs ou namespaces NVMe) são criadas a partir de uma camada de armazenamento simplificada chamada Zonas de Disponibilidade de Armazenamento (SAZs),

que atuam como pools comuns de armazenamento para um par de alta disponibilidade (HA).

Normalmente, existe apenas uma zona de disponibilidade de armazenamento (SAZ) por par de alta disponibilidade (HA).

Zonas de Disponibilidade de Armazenamento (SAZ)

Nos sistemas ASA r2, os volumes ainda existem, mas são criados automaticamente quando as unidades de armazenamento são criadas. As unidades de armazenamento (LUNs ou namespaces NVMe) são provisionadas diretamente nos volumes criados automaticamente nas Zonas de Disponibilidade de Armazenamento (SAZs). Este design elimina a necessidade de gerenciamento manual de volumes e torna o provisionamento mais direto e simplificado para cargas de trabalho em bloco, como bancos de dados Oracle.

Zonas de Ação Especial (ZAE) e unidades de armazenamento

As unidades de armazenamento relacionadas (LUNs ou namespaces NVMe) normalmente estão localizadas no mesmo espaço de armazenamento, dentro de uma única Zona de Disponibilidade de Armazenamento (SAZ). Por exemplo, um banco de dados que requer 10 unidades de armazenamento (LUNs) normalmente teria todas as 10 unidades colocadas na mesma SAZ (Zona de Armazenamento de Dados) para maior simplicidade e melhor desempenho.

- O comportamento padrão do ASA r2 é usar uma proporção de 1:1 entre unidades de armazenamento e volumes, ou seja, uma unidade de armazenamento (LUN) por volume.
- Caso haja mais de um par HA no sistema ASA r2, as unidades de armazenamento (LUNs) para um determinado banco de dados podem ser distribuídas por várias SAZs para otimizar a utilização e o desempenho do controlador.

No contexto de FC SAN, a unidade de armazenamento se refere a LUN.

Grupos de consistência (CGs), LUNs e snapshots

No ASA r2, as políticas e agendamentos de snapshots são aplicados no nível do Grupo de Consistência, que é uma construção lógica que agrupa vários LUNs ou namespaces NVMe para proteção de dados coordenada. Um conjunto de dados composto por 10 LUNs exigiria apenas uma única política de snapshot quando esses LUNs fizessem parte do mesmo Grupo de Consistência.

Os Grupos de Consistência garantem operações de snapshot atômicas em todos os LUNs incluídos. Por exemplo, um banco de dados que reside em 10 LUNs, ou um ambiente de aplicativos baseado em VMware composto por 10 sistemas operacionais diferentes, pode ser protegido como um único objeto consistente se os LUNs subjacentes estiverem agrupados no mesmo grupo de consistência. Se forem colocados em grupos de consistência diferentes, os snapshots podem ou não estar perfeitamente sincronizados, mesmo que agendados para o mesmo horário.

Em alguns casos, um conjunto relacionado de LUNs pode precisar ser dividido em dois grupos de consistência diferentes devido a requisitos de recuperação. Por exemplo, um banco de dados pode ter quatro LUNs para arquivos de dados e dois LUNs para logs. Nesse caso, um grupo de consistência de arquivos de dados com 4 LUNs e um grupo de consistência de logs com 2 LUNs podem ser a melhor opção. O motivo é a recuperabilidade independente: o grupo de consistência do arquivo de dados poderia ser restaurado seletivamente para um estado anterior, o que significa que todos os quatro LUNs seriam revertidos para o estado do snapshot, enquanto o grupo de consistência do log, com seus dados críticos, permaneceria intacto.

CGs, LUNs e SnapMirror

As políticas e operações do SnapMirror , assim como as operações de snapshot, são executadas no grupo de consistência, e não no LUN.

A localização conjunta de LUNs relacionadas em um único grupo de consistência permite criar um único relacionamento SnapMirror e atualizar todos os dados contidos nele com uma única atualização. Assim como nos snapshots, a atualização também será uma operação atômica. O destino SnapMirror teria a garantia de possuir uma réplica pontual dos LUNs de origem. Se os LUNs estiverem distribuídos por vários grupos de consistência, as réplicas podem ou não ser consistentes entre si.

A replicação SnapMirror em sistemas ASA r2 apresenta as seguintes limitações:

- A replicação síncrona do SnapMirror não é suportada.
- O SnapMirror ActiveSync é compatível apenas entre dois sistemas ASA R2.
- A replicação assíncrona do SnapMirror é suportada apenas entre dois sistemas ASA r2.
- A replicação assíncrona do SnapMirror não é suportada entre um sistema ASA r2 e um sistema ASA, AFF ou FAS , ou a nuvem.

Saiba mais sobre "Políticas de replicação SnapMirror suportadas em sistemas ASA r2".

CGs, LUNs e QoS

Embora o QoS possa ser aplicado seletivamente a LUNs individuais, geralmente é mais fácil configurá-lo no nível do grupo de consistência. Por exemplo, todos os LUNs usados pelos sistemas operacionais convidados em um determinado servidor ESX podem ser colocados em um único grupo de consistência e, em seguida, uma política de QoS adaptativa do ONTAP pode ser aplicada. O resultado é um limite de IOPS por TiB com escalonamento automático que se aplica a todos os LUNs.

Da mesma forma, se um banco de dados exigisse 100 mil IOPS e ocupasse 10 LUNs, seria mais fácil definir um único limite de 100 mil IOPS em um único grupo de consistência do que definir 10 limites individuais de 10 mil IOPS, um em cada LUN.

Vários layouts CG

Há casos em que distribuir LUNs por vários grupos de consistência pode ser benéfico. O principal motivo é a distribuição de dados entre os controladores. Por exemplo, um sistema de armazenamento HA ASA r2 pode hospedar um único banco de dados Oracle, onde todo o potencial de processamento e cache de cada controlador é necessário. Nesse caso, um projeto típico seria colocar metade dos LUNs em um único grupo de consistência no controlador 1 e a outra metade dos LUNs em um único grupo de consistência no controlador 2.

Da mesma forma, em ambientes que hospedam muitos bancos de dados, a distribuição de LUNs em vários grupos de consistência pode garantir uma utilização equilibrada do controlador. Por exemplo, um sistema de alta disponibilidade (HA) que hospeda 100 bancos de dados com 10 LUNs cada pode atribuir 5 LUNs a um grupo de consistência no controlador 1 e 5 LUNs a um grupo de consistência no controlador 2 por banco de dados. Isso garante um carregamento simétrico à medida que bancos de dados adicionais são provisionados.

Nenhum desses exemplos envolve uma proporção de 1:1 entre LUN e grupo de consistência. O objetivo continua sendo otimizar a capacidade de gerenciamento agrupando LUNs relacionados logicamente em um grupo de consistência.

Um exemplo em que uma proporção de 1:1 entre LUN e grupo de consistência faz sentido são as cargas de trabalho conteinerizadas, onde cada LUN pode representar uma única carga de trabalho que requer políticas

de snapshot e replicação separadas e, portanto, precisa ser gerenciada individualmente. Nesses casos, uma proporção de 1:1 pode ser ideal.

Redimensionamento LUN e redimensionamento LVM

Quando um sistema de arquivos baseado em SAN ou um grupo de discos Oracle ASM atinge seu limite de capacidade no ASA r2, existem duas opções para aumentar o espaço disponível:

- Aumentar o tamanho das LUNs (unidades de armazenamento) existentes.
- Adicione um novo LUN a um grupo de discos ASM ou grupo de volumes LVM existente e expanda o volume lógico contido.

Embora o redimensionamento de LUN seja suportado no ASA r2, geralmente é melhor usar um Gerenciador de Volumes Lógicos (LVM), como o Oracle ASM. Uma das principais razões para a existência dos LVMs é evitar a necessidade de redimensionamento frequente de LUNs. Com um LVM, vários LUNs são combinados em um pool virtual de armazenamento. Os volumes lógicos criados a partir desse pool podem ser facilmente redimensionados sem afetar a configuração de armazenamento subjacente.

Outros benefícios do uso de LVM ou ASM incluem:

- Otimização de desempenho: Distribui as operações de E/S por vários LUNs, reduzindo os pontos de acesso intenso.
- Flexibilidade: Adicione novos LUNs sem interromper as cargas de trabalho existentes.
- Migração transparente: ASM ou LVM podem realocar extensões para novos LUNs para balanceamento ou tiering sem tempo de inatividade do host.

> Principais considerações sobre a ASA r2:
>
> - O redimensionamento de LUN é realizado no nível da unidade de armazenamento dentro de uma Máquina Virtual de Armazenamento (SVM), utilizando a capacidade da Zona de Disponibilidade de Armazenamento (SAZ).
> - Para a Oracle, a melhor prática é adicionar LUNs aos grupos de discos ASM em vez de redimensionar os LUNs existentes, para manter o striping e o paralelismo.

LVM striping (distribuição LVM)

A distribuição de LVM refere-se à distribuição de dados entre vários LUNs. O resultado é uma melhoria significativa na performance de muitos bancos de dados.

Antes da era das unidades flash, a distribuição foi usada para ajudar a superar as limitações de desempenho das unidades giratórias. Por exemplo, se um sistema operacional precisar executar uma operação de leitura 1MB, ler que 1MB TB de dados de uma única unidade exigiria muita busca e leitura de cabeça de unidade, pois o 1MB é transferido lentamente. Se esse 1MB TB de dados fosse distribuído em 8 LUNs, o sistema operacional poderia emitir oito operações de leitura 128K em paralelo e reduzir o tempo necessário para concluir a transferência 1MB.

A gravação em faixas com discos rígidos era mais difícil porque o padrão de E/S precisava ser conhecido antecipadamente. Se o striping não estiver configurado corretamente para os padrões reais de E/S, as configurações em striping podem prejudicar o desempenho. Com bancos de dados Oracle, e especialmente com configurações de armazenamento totalmente em flash, o striping é muito mais fácil de configurar e

comprovadamente melhora drasticamente o desempenho.

Gerenciadores de volume lógicos, como o Oracle ASM stripe por padrão, mas o LVM de SO nativo não. Alguns deles unem vários LUNs como um dispositivo concatenado, o que resulta em datafiles que existem em um e apenas um dispositivo LUN. Isso causa pontos quentes. Outras implementações de LVM são padrão para extensões distribuídas. Isso é semelhante ao striping, mas é mais grosseiro. Os LUNs no grupo de volumes são cortados em pedaços grandes, chamados de extensões e normalmente medidos em muitos megabytes, e os volumes lógicos são então distribuídos por essas extensões. O resultado é a e/S aleatória contra um arquivo deve ser bem distribuída entre LUNs, mas as operações de e/S sequenciais não são tão eficientes quanto poderiam ser.

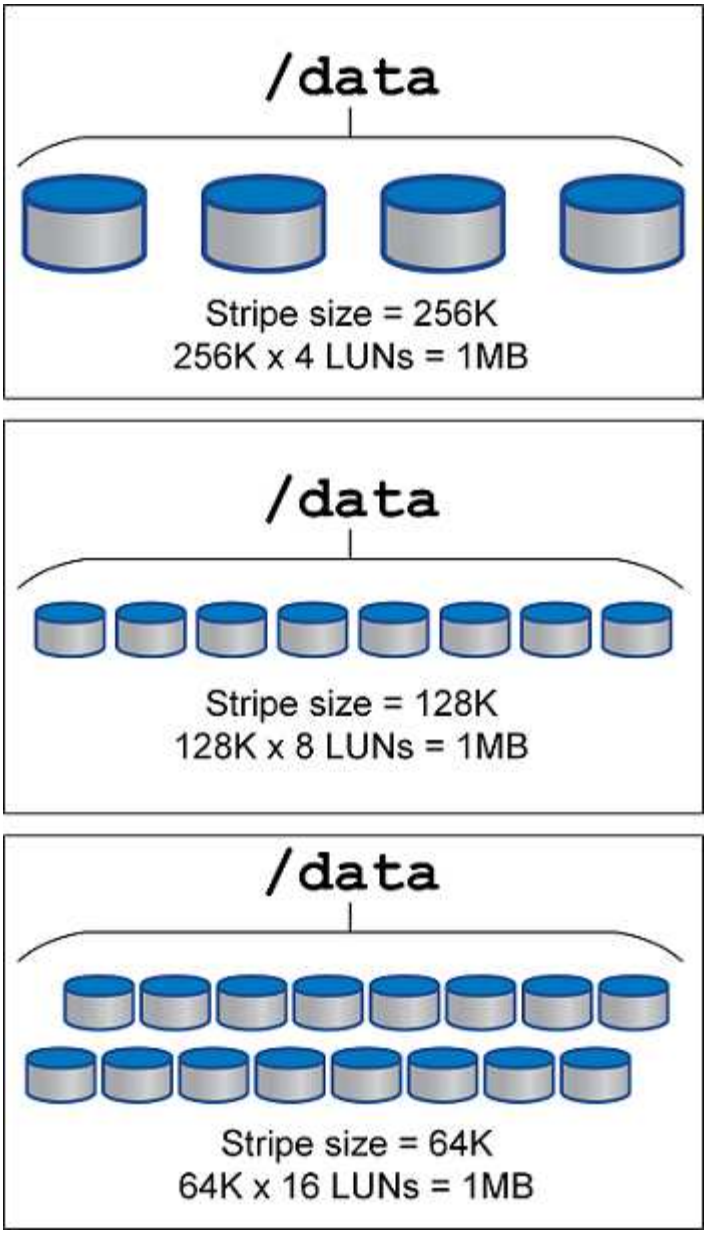
A e/S de aplicativos intensivos quase sempre é (a) em unidades do tamanho básico do bloco ou (b) um megabyte.

O principal objetivo de uma configuração de distribuição é garantir que a e/S de arquivo único possa ser executada como uma única unidade, e as e/S de vários blocos, que devem ter 1MB MB de tamanho, possam ser paralelizadas uniformemente em todos os LUNs no volume distribuído. Isso significa que o tamanho do stripe não deve ser menor que o tamanho do bloco do banco de dados e o tamanho do stripe multiplicado pelo número de LUNs deve ser 1MB.

Melhores práticas para o uso de LVM striping em bancos de dados Oracle:

- Tamanho da faixa ≥ tamanho do bloco do banco de dados.
- Tamanho da faixa * número de LUNs ≈ 1 MB para paralelismo ideal.
- Utilize vários LUNs por grupo de discos ASM para maximizar a taxa de transferência e evitar pontos de acesso intenso.

A figura a seguir mostra três opções possíveis para o tamanho da faixa e ajuste de largura. O número de LUNs é selecionado para atender aos requisitos de performance conforme descrito acima, mas, em todos os casos, o total de dados em um único stripe é de 1MB.



NVFAIL

NVFAIL é um recurso do ONTAP que garante a integridade dos dados durante cenários de falha catastrófica.

Essa funcionalidade ainda é aplicável em sistemas ASA r2, mesmo que o ASA r2 utilize uma arquitetura SAN simplificada (SAZs e unidades de armazenamento em vez de volumes).

Os bancos de dados são vulneráveis à corrupção durante eventos de falha de armazenamento porque mantêm grandes caches internos. Se um evento catastrófico exigir a reinicialização forçada do ONTAP , independentemente do estado de conservação da configuração geral, as alterações previamente reconhecidas poderão ser descartadas. O conteúdo da matriz de armazenamento retrocede no tempo, e o estado do cache do banco de dados deixa de refletir o estado dos dados no disco. Essa inconsistência resulta em corrupção de dados.

O armazenamento em cache pode ocorrer na camada de aplicação ou na camada de servidor. Por exemplo, uma configuração do Oracle Real Application Cluster (RAC) com servidores ativos tanto em um site primário

quanto em um site remoto armazena dados em cache dentro da SGA do Oracle. Uma operação de failover forçada que resultasse em perda de dados colocaria o banco de dados em risco de corrupção, pois os blocos armazenados na SGA poderiam não corresponder aos blocos no disco.

Um uso menos óbvio do cache ocorre na camada do sistema de arquivos do sistema operacional. Um sistema de arquivos em cluster baseado em LUNs localizados no site primário poderia ser montado em servidores no site remoto, e mais uma vez os dados poderiam ser armazenados em cache. Uma falha na NVRAM ou uma tomada de controle forçada nessas situações pode resultar em corrupção do sistema de arquivos.

O ONTAP protege bancos de dados e sistemas operacionais contra esse cenário usando o NVFAIL e suas configurações associadas, que sinalizam ao host para invalidar os dados em cache e remontar os sistemas de arquivos afetados após o failover. Esse mecanismo se aplica aos LUNs e namespaces do ASA r2 da mesma forma que no AFF/ FAS.

Principais considerações sobre a ASA r2:

- O NVFAIL opera no nível do LUN (unidade de armazenamento), não no nível do SAZ.
- Para bancos de dados Oracle, o NVFAIL deve ser habilitado em todos os LUNs que hospedam componentes críticos (arquivos de dados, logs de refazer, arquivos de controle).
- O MetroCluster não é suportado no ASA r2, portanto, o NVFAIL se aplica principalmente a cenários de failover de alta disponibilidade local.
- O NFS não é suportado no ASA r2, portanto, as considerações sobre NVFAIL aplicam-se apenas a cargas de trabalho baseadas em SAN (FC/iSCSI/NVMe).

Utilitário de recuperação ASM (ASRU)

O ONTAP no ASA r2 remove de forma eficiente os blocos zerados gravados em um LUN (unidade de armazenamento) quando a compressão inline está habilitada. Utilitários como o Oracle ASM Reclamation Utility (ASRU) funcionam gravando zeros em extensões ASM não utilizadas.

Isso permite que os administradores de banco de dados recuperem espaço no array de armazenamento após a exclusão de dados. O ONTAP intercepta os zeros e desaloca o espaço do LUN. O processo de recuperação é extremamente rápido porque nenhum dado real é gravado no sistema de armazenamento.

Do ponto de vista do banco de dados, o grupo de discos ASM contém zeros, e a leitura dessas regiões dos LUNs resultaria em um fluxo de zeros, mas o ONTAP não armazena os zeros nas unidades. Em vez disso, são feitas mudanças simples de metadados que marcam internamente as regiões zeroadas do LUN como vazias de qualquer dado.

Por motivos semelhantes, os testes de desempenho envolvendo dados isolados não são válidos, pois blocos de zeros não são realmente processados como gravações no storage array.

Principais considerações sobre ASRU com ASA r2 ONTAP:

- Funciona da mesma forma que o AFF/ FAS para cargas de trabalho SAN, pois o ASA r2 é somente para blocos.
- Aplica-se a LUNs e namespaces NVMe provisionados em SAZs.
- Não existem volumes FlexVol , mas o comportamento de recuperação de blocos zerados é idêntico.

Ao usar o ASRU, certifique-se de que todos os patches recomendados pela Oracle estejam instalados.

Virtualização

A virtualização de bancos de dados com VMware, Oracle OLVM ou KVM é uma escolha cada vez mais comum para os clientes da NetApp que escolheram a virtualização até mesmo para seus bancos de dados mais críticos.

Capacidade de suporte

Existem muitos equívocos sobre as políticas de suporte Oracle para virtualização, especialmente para produtos VMware. Não é incomum ouvir que o Oracle outright não suporta virtualização. Essa noção está incorreta e leva a oportunidades perdidas para se beneficiar da virtualização. O Oracle Doc ID 249212,1 discute os requisitos reais e raramente é considerado pelos clientes como uma preocupação.

Se ocorrer um problema em um servidor virtualizado e esse problema for anteriormente desconhecido para o suporte Oracle, o cliente poderá ser solicitado a reproduzir o problema em hardware físico. Um cliente Oracle que executa uma versão de ponta a ponta de um produto pode não querer usar a virtualização devido ao potencial de problemas de capacidade de suporte, mas essa situação não tem sido um mundo real para clientes de virtualização que usam versões de produtos Oracle geralmente disponíveis.

Apresentação de armazenamento

Os clientes que consideram a virtualização de seus bancos de dados devem basear suas decisões de storage em suas necessidades de negócios. Embora esta seja uma declaração geralmente verdadeira para todas as DECISÕES DE TI, é especialmente importante para projetos de banco de dados, porque o tamanho e o escopo dos requisitos variam consideravelmente.

Existem três opções básicas para apresentação de armazenamento:

- LUNs virtualizados em datastores do hipervisor
- ISCSI LUNs gerenciados pelo iniciador iSCSI na VM, não pelo hipervisor
- Sistemas de arquivos NFS montados pela VM (não em um datastore baseado em NFS)
- Mapeamentos diretos de dispositivos. Os RDMs VMware são desfavorecidos pelos clientes, mas os dispositivos físicos ainda são frequentemente mapeados diretamente com a virtualização KVM e OLVM.

Desempenho

O método de apresentar armazenamento a um convidado virtualizado geralmente não afeta o desempenho. Os sistemas operacionais de host, drivers de rede virtualizados e implementações de armazenamento de dados do hipervisor são altamente otimizados e geralmente podem consumir toda a largura de banda de rede FC ou IP disponível entre o hipervisor e o sistema de storage, desde que sejam seguidas as práticas recomendadas básicas. Em alguns casos, obter desempenho ideal pode ser um pouco mais fácil usando uma abordagem de apresentação de storage em comparação com outra, mas o resultado final deve ser comparável.

Capacidade de gerenciamento

O fator chave para decidir como apresentar o storage a um convidado virtualizado é a capacidade de gerenciamento. Não há método certo ou errado. A melhor abordagem depende das necessidades

operacionais, habilidades e preferências DE TI.

Os fatores a considerar incluem:

- **Transparência.** Quando uma VM gerencia seus sistemas de arquivos, é mais fácil para um administrador de banco de dados ou um administrador de sistema identificar a origem dos sistemas de arquivos para seus dados. Os sistemas de arquivos e LUNs são acessados de forma diferente do que com um servidor físico.
- **Consistência.** Quando uma VM possui seus sistemas de arquivos, o uso ou não uso de uma camada de hipervisor afeta a gerenciabilidade. Os mesmos procedimentos de provisionamento, monitoramento, proteção de dados etc. podem ser usados em todo o estado, incluindo ambientes virtualizados e não virtualizados.

  Por outro lado, em um data center 100% virtualizado, pode ser preferível também usar o armazenamento baseado em datastore em toda a área ocupada com a mesma lógica mencionada acima - consistência - a capacidade de usar os mesmos procedimentos para provisionamento, proteção, monitoramento e proteção de dados.

- **Estabilidade e resolução de problemas.** Quando uma VM é proprietária de seus sistemas de arquivos, fornecer performance boa e estável e solucionar problemas fica mais simples porque toda a pilha de storage está presente na VM. A única função do hipervisor é transportar quadros FC ou IP. Quando um datastore é incluído em uma configuração, complica a configuração introduzindo outro conjunto de timeouts, parâmetros, arquivos de log e possíveis bugs.
- * Portabilidade.* Quando uma VM possui seus sistemas de arquivos, o processo de mover um ambiente Oracle se torna muito mais simples. Os sistemas de arquivos podem ser movidos facilmente entre convidados virtualizados e não virtualizados.
- **Aprisionamento do fornecedor.** depois que os dados são colocados em um datastore, usar um hypervisor diferente ou tirar os dados do ambiente virtualizado torna-se completamente difícil.
- **Capacitação de snapshot.** Os procedimentos tradicionais de backup em um ambiente virtualizado podem se tornar um problema devido à largura de banda relativamente limitada. Por exemplo, um tronco 10GbE de quatro portas pode ser suficiente para suportar as necessidades diárias de desempenho de muitos bancos de dados virtualizados, mas esse tronco seria insuficiente para executar backups usando RMAN ou outros produtos de backup que exigem o streaming de uma cópia de tamanho completo dos dados. O resultado é que um ambiente virtualizado cada vez mais consolidado precisa executar backups por meio de snapshots de storage. Isso evita a necessidade de sobrecompilar a configuração do hipervisor apenas para suportar os requisitos de largura de banda e CPU na janela de backup.

  O uso de sistemas de arquivos de propriedade de convidados às vezes facilita a utilização de backups e restaurações baseados em snapshot, porque os objetos de storage que precisam de proteção podem ser segmentados com mais facilidade. No entanto, há um número cada vez maior de produtos de proteção de dados de virtualização que se integram bem aos armazenamentos de dados e snapshots. A estratégia de backup deve ser totalmente considerada antes de tomar uma decisão sobre como apresentar o storage a um host virtualizado.

Drivers paravirtualizados

Para um desempenho ideal, o uso de drivers de rede paravirtualizados é fundamental. Quando um datastore é usado, um driver SCSI paravirtualizado é necessário. Um driver de dispositivo paravirtualizado permite que um convidado se integre mais profundamente no hypervisor, em vez de um driver emulado no qual o hypervisor passa mais tempo de CPU imitando o comportamento do hardware físico.

Sobrecarga da RAM

Overcommit RAM significa configurar mais RAM virtualizada em vários hosts do que existe no hardware físico. Isso pode causar problemas inesperados de desempenho. Ao virtualizar um banco de dados, os blocos subjacentes do Oracle SGA não devem ser trocados pelo hipervisor para storage. Isso causa resultados de desempenho altamente instáveis.

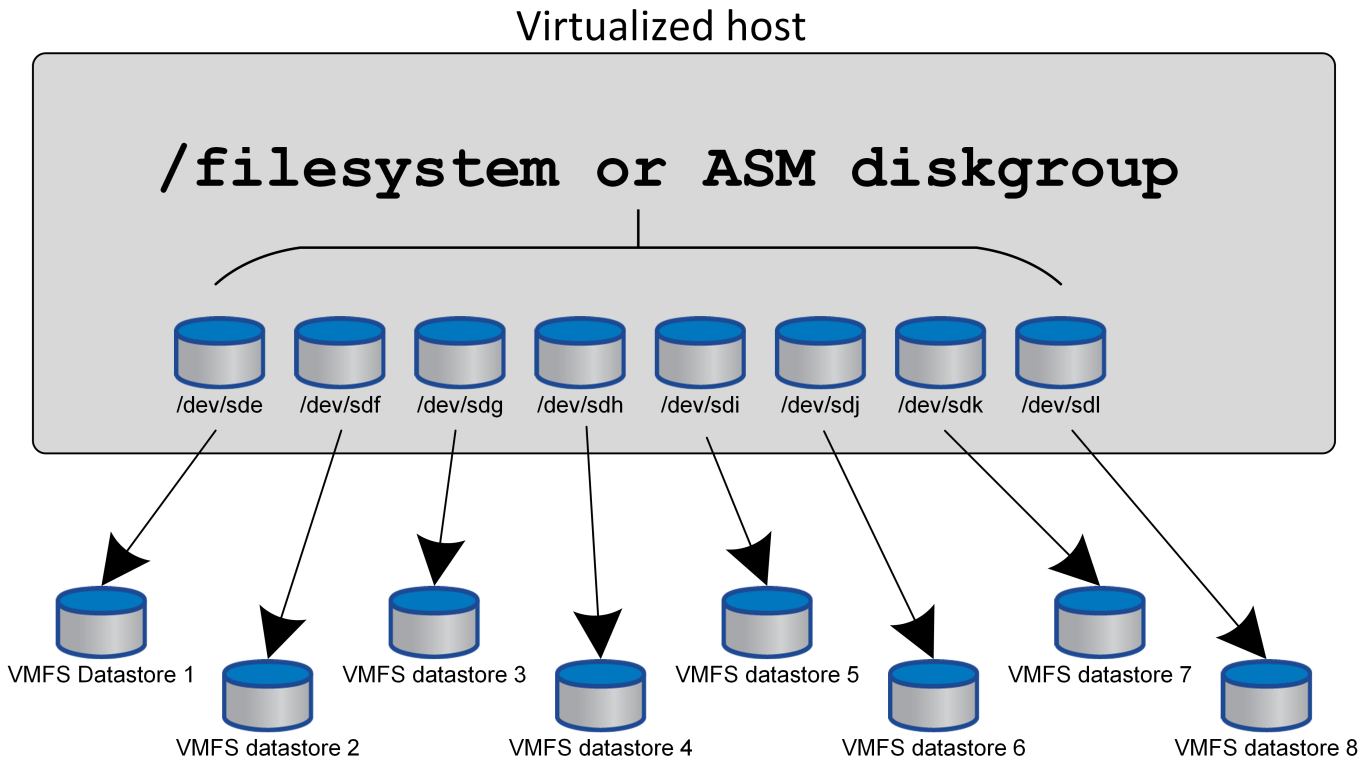
Striping do datastore

Ao usar bancos de dados com datastores, há um fator crítico a considerar em relação ao desempenho - striping.

As tecnologias de armazenamento de dados, como o VMFS, são capazes de abranger vários LUNs, mas não são dispositivos distribuídos. Os LUNs são concatenados. O resultado final pode ser pontos quentes LUN. Por exemplo, um banco de dados Oracle típico pode ter um grupo de discos ASM de 8 LUN. Todos os 8 LUNs virtualizados podem ser provisionados em um datastore VMFS de 8 LUN, mas não há garantia sobre quais LUNs os dados residirão. A configuração resultante pode ser todo 8 LUN virtualizado ocupando um único LUN dentro do datastore VMFS. Isso se torna um gargalo de desempenho.

Geralmente é necessário riscar. Com alguns hypervisors, incluindo KVM, é possível criar um datastore usando o striping LVM como descrito "aqui". Com a VMware, a arquitetura parece um pouco diferente. Cada LUN virtualizado precisa ser colocado em um datastore VMFS diferente.

Por exemplo:



O driver principal para essa abordagem não é o ONTAP, é devido à limitação inerente do número de operações que uma única VM ou LUN de hipervisor pode servir em paralelo. Um único LUN de ONTAP geralmente suporta muito mais IOPS do que um host pode solicitar. O limite de desempenho de LUN único é quase universalmente um resultado do sistema operacional do host. Assim, a maioria dos bancos de dados precisa de 4 a 8 LUNs para atender às suas necessidades de performance.

As arquiteturas VMware precisam Planejar suas arquiteturas cuidadosamente para garantir que as máximas de armazenamento de dados e/ou caminho LUN não sejam encontradas com essa abordagem. Addtionally, não há nenhum requisito para um conjunto exclusivo de armazenamentos de dados VMFS para cada banco de dados. A principal necessidade é garantir que cada host tenha um conjunto limpo de 4-8 caminhos de e/S, desde os LUNs virtualizados até os LUNs de back-end no próprio sistema de storage. Em raras ocasiões, ainda mais datadores podem ser benéficos para demandas de desempenho verdadeiramente extremas, mas 4-8 LUNs geralmente são suficientes para 95% de todos os bancos de dados. Um único volume ONTAP contendo 8 LUNs pode suportar até 250.000 IOPS de bloco Oracle aleatório com uma configuração típica de os/ONTAP/rede.

Disposição em camadas

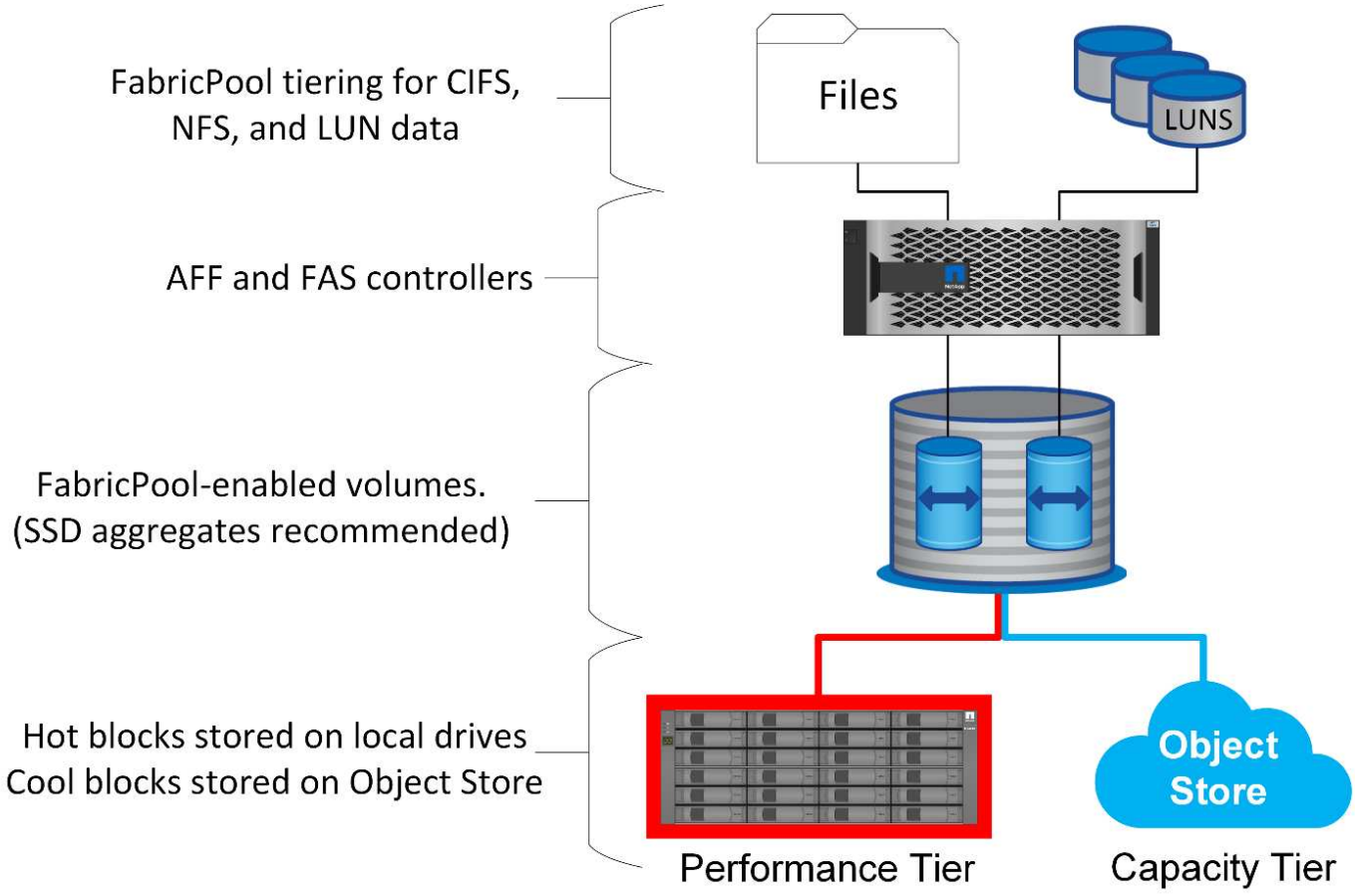
Visão geral

Entender como a disposição em camadas do FabricPool afeta a Oracle e outros bancos de dados requer um entendimento da arquitetura FabricPool de baixo nível.

Arquitetura

O FabricPool é uma tecnologia em camadas que classifica os blocos como ativos ou inativos e os coloca na camada mais apropriada de storage. A camada de performance fica na maioria das vezes localizada no storage SSD e hospeda os blocos de dados ativos. A camada de capacidade fica em um armazenamento de objetos e hospeda os blocos de dados inativos. O suporte ao storage de objetos inclui NetApp StorageGRID, ONTAP S3, storage Microsoft Azure Blob, serviço de storage de objetos Alibaba Cloud, IBM Cloud Object Storage, storage do Google Cloud e Amazon AWS S3.

Várias políticas de disposição em camadas estão disponíveis para controlar a classificação dos blocos como ativo ou inativo. Além disso, as políticas podem ser definidas por volume e alteradas conforme necessário. Somente os blocos de dados são movidos entre as categorias de performance e capacidade. Os metadados que definem a estrutura do sistema de arquivos e LUN sempre permanecem no nível de performance. Como resultado, o gerenciamento é centralizado no ONTAP. Os arquivos e LUNs não são diferentes dos dados armazenados em qualquer outra configuração do ONTAP. O controlador NetApp AFF ou FAS aplica as políticas definidas para mover dados para o nível apropriado.



Provedores de armazenamento de objetos

Os protocolos de armazenamento de objetos usam solicitações HTTP ou HTTPS simples para armazenar um grande número de objetos de dados. O acesso ao armazenamento de objetos deve ser confiável, pois o acesso aos dados do ONTAP depende do atendimento imediato das solicitações. As opções incluem as opções de acesso padrão e pouco frequentes do Amazon S3, além do Microsoft Azure Hot and Cool Blob Storage, IBM Cloud e Google Cloud. As opções de arquivamento, como o Amazon Glacier e o Amazon Archive, não são suportadas porque o tempo necessário para recuperar dados pode exceder as tolerâncias dos sistemas operacionais e aplicativos host.

O NetApp StorageGRID também é compatível e é uma solução de classe empresarial ideal. Ele é um sistema de storage de objetos de alto desempenho, dimensionável e altamente seguro que pode fornecer redundância geográfica para dados do FabricPool, bem como outras aplicações de armazenamento de objetos que provavelmente farão parte dos ambientes de aplicativos empresariais.

O StorageGRID também pode reduzir custos evitando as cobranças de saída impostas por muitos fornecedores de nuvem pública pela leitura de dados de seus serviços.

Dados e metadados

Observe que o termo "dados" aqui se aplica aos blocos de dados reais, não aos metadados. Apenas os blocos de dados são dispostos em camadas, enquanto os metadados permanecem na camada de performance. Além disso, o status de um bloco como quente ou frio só é afetado pela leitura do bloco de dados real. A simples leitura do nome, carimbo de data/hora ou metadados de propriedade de um arquivo não afeta a localização dos blocos de dados subjacentes.

Backups

Embora o FabricPool possa reduzir significativamente o espaço físico do storage, ele não é por si só uma solução de backup. Os metadados do NetApp WAFL sempre permanecem no nível de performance. Se um desastre catastrófico destruir o nível de performance, não será possível criar um novo ambiente usando os dados na categoria de capacidade porque não contém metadados do WAFL.

No entanto, o FabricPool pode se tornar parte de uma estratégia de backup. Por exemplo, o FabricPool pode ser configurado com a tecnologia de replicação NetApp SnapMirror. Cada metade do espelho pode ter sua própria conexão com um destino de armazenamento de objetos. O resultado são duas cópias independentes dos dados. A cópia primária consiste em blocos na camada de performance e blocos associados na camada de capacidade, e a réplica é um segundo conjunto de blocos de performance e capacidade.

Políticas de disposição em camadas

Políticas de disposição em camadas

Quatro políticas estão disponíveis no ONTAP, que controlam como os dados da Oracle no nível de desempenho se tornam um candidato a ser relocado para o nível de capacidade.

Apenas Snapshot

O `snapshot-only` `tiering-policy` aplica-se apenas a blocos que não são partilhados com o sistema de ficheiros ativo. Essencialmente, isso resulta em camadas de backups de bancos de dados. Os blocos se tornam candidatos à disposição em camadas depois que um snapshot é criado e o bloco é substituído, resultando em um bloco que existe apenas no snapshot. O atraso antes de um `snapshot-only` bloco ser considerado frio é controlado pela `tiering-minimum-cooling-days` definição do volume. A gama a partir de ONTAP 9.8 é de 2 a 183 dias.

Muitos conjuntos de dados têm taxas de alteração baixas, o que resulta em economias mínimas com essa política. Por exemplo, um banco de dados típico observado no ONTAP tem uma taxa de alteração inferior a 5% por semana. Os logs de arquivo de banco de dados podem ocupar um espaço extenso, mas eles geralmente continuam a existir no sistema de arquivos ativo e, portanto, não seriam candidatos para a disposição em camadas sob esta política.

Auto

A `auto` política de disposição em camadas estende a disposição em camadas para blocos específicos de snapshot e blocos no sistema de arquivos ativo. O atraso antes de um bloco ser considerado frio é controlado pela `tiering-minimum-cooling-days` definição do volume. A gama a partir de ONTAP 9.8 é de 2 a 183 dias.

Essa abordagem permite opções de disposição em categorias que não estão disponíveis na `snapshot-only` política. Por exemplo, uma política de proteção de dados pode exigir 90 dias de retenção de determinados arquivos de log. Definir um período de resfriamento de 3 dias resulta em todos os arquivos de log com mais de 3 dias para serem dispostos na camada de desempenho. Essa ação libera espaço substancial na categoria de performance e ainda permite que você visualize e gerencie todos os 90 dias de dados.

Nenhum

A `none` política de disposição em categorias impede que blocos adicionais sejam dispostos em camadas na camada de storage, mas todos os dados ainda na camada de capacidade permanecem na camada de capacidade até que sejam lidos. Se o bloco for então lido, ele será puxado para trás e colocado no nível de

desempenho.

O principal motivo para usar a `none` política de disposição em camadas é impedir que blocos sejam dispostos em camadas, mas pode se tornar útil alterar as políticas ao longo do tempo. Por exemplo, digamos que um conjunto de dados específico seja amplamente categorizado na camada de capacidade, mas surge uma necessidade inesperada de recursos completos de performance. A política pode ser alterada para impedir qualquer disposição em camadas adicional e para confirmar que todos os blocos de leitura de volta à medida que a IO aumenta permanecem na categoria de performance.

Tudo

A `all` política de disposição em categorias substitui a `backup` política a partir do ONTAP 9.6. `backup`A política aplicada somente a volumes de proteção de dados, o que significa um destino do SnapMirror ou do NetApp SnapVault. A `all` política funciona da mesma forma, mas não se restringe a volumes de proteção de dados.

Com essa política, os blocos são imediatamente considerados frios e elegíveis para serem dispostos na camada de capacidade imediatamente.

Essa política é especialmente apropriada para backups de longo prazo. Ele também pode ser usado como uma forma de Gerenciamento de armazenamento hierárquico (HSM). No passado, o HSM era comumente usado para categorizar os blocos de dados de um arquivo em fita, mantendo o próprio arquivo visível no sistema de arquivos. Um volume FabricPool com a `all` política permite armazenar arquivos em um ambiente visível e gerenciável, mas não consome praticamente espaço na camada de storage local.

Políticas de recuperação

As políticas de disposição em camadas controlam quais blocos de banco de dados Oracle são dispostos da camada de performance para a camada de capacidade. As políticas de recuperação controlam o que acontece quando um bloco que foi categorizado é lido.

Padrão

Todos os volumes do FabricPool são inicialmente definidos como `default`, o que significa que o comportamento é controlado pela política de recuperação de nuvem. "O comportamento exato depende da política de disposição em camadas usada.

- `auto`- apenas recuperar dados lidos aleatoriamente
- `snapshot-only`- recuperar todos os dados de leitura sequencial ou aleatória
- `none`- recuperar todos os dados de leitura sequencial ou aleatória
- `all`- não recupere dados do nível de capacidade

Na leitura

A configuração `cloud-retrieval-policy` como On-read substitui o comportamento padrão para que uma leitura de qualquer dado em camadas resulte no retorno desses dados ao nível de desempenho.

Por exemplo, um volume pode ter sido levemente usado por um longo tempo sob `auto` a política de disposição em camadas e a maioria dos blocos agora está dividida.

Se uma mudança inesperada nas necessidades de negócios exigir que alguns dos dados sejam verificados

repetidamente para preparar um determinado relatório, pode ser desejável alterar o `cloud-retrieval-policy` para para `on-read` para garantir que todos os dados lidos sejam devolvidos ao nível de desempenho, incluindo dados de leitura sequencial e aleatória. Isso melhoraria o desempenho de e/S sequenciais em relação ao volume.

Promover

O comportamento da política promover depende da política de disposição em camadas. Se a política de disposição em categorias for `auto`, a configuração do `cloud-retrieval-policy `to `promote` recuperará todos os blocos da camada de capacidade na próxima verificação de disposição em categorias.

Se a política de disposição em categorias for `snapshot-only`, os únicos blocos retornados serão os blocos associados ao sistema de arquivos ativo. Normalmente, isso não teria nenhum efeito porque os únicos blocos dispostos sob `snapshot-only` a política seriam blocos associados exclusivamente a snapshots. Não haveria blocos em camadas no sistema de arquivos ativo.

No entanto, se os dados em um volume tiverem sido restaurados por uma operação de clone de arquivo ou SnapRestore de volume a partir de um snapshot, alguns dos blocos que foram dispostos em camadas por estarem associados apenas a snapshots podem agora ser exigidos pelo sistema de arquivos ativo. Pode ser desejável alterar temporariamente a `cloud-retrieval-policy` política para `promote` recuperar rapidamente todos os blocos necessários localmente.

Nunca

Não recupere blocos da camada de capacidade.

Estratégias de disposição em camadas

Disposição completa de arquivos em categorias

Embora a disposição em camadas do FabricPool opere no nível de bloco, em alguns casos, ela pode ser usada para fornecer a disposição em camadas no nível do arquivo.

Muitos conjuntos de dados de aplicações são organizados por data, e esses dados geralmente têm cada vez menos probabilidade de serem acessados à medida que envelhecem. Por exemplo, um banco pode ter um repositório de arquivos PDF que contém cinco anos de declarações de clientes, mas apenas os últimos meses estão ativos. O FabricPool pode ser usado para realocar arquivos de dados mais antigos para o nível de capacidade. Um período de resfriamento de 14 dias garantiria que os 14 dias mais recentes de arquivos PDF permaneçam no nível de desempenho. Além disso, os arquivos que são lidos pelo menos a cada 14 dias permanecerão ativos e, portanto, permanecerão no nível de desempenho.

Políticas

Para implementar uma abordagem de disposição em camadas baseada em arquivos, você precisa ter arquivos gravados e não modificados posteriormente. A `tiering-minimum-cooling-days` política deve ser definida suficientemente alta para que os arquivos que você possa precisar permaneçam no nível de desempenho. Por exemplo, um conjunto de dados para o qual os 60 dias de dados mais recentes são necessários com garantias de desempenho ideais, definindo `tiering-minimum-cooling-days` o período como 60. Resultados semelhantes também podem ser alcançados com base nos padrões de acesso ao arquivo. Por exemplo, se os 90 dias mais recentes de dados forem necessários e o aplicativo estiver acessando esse período de 90 dias de dados, os dados permanecerão na categoria de performance. Ao definir `tiering-minimum-cooling-days` o período como 2, você obtém a disposição imediata de categorias após os dados ficarem menos ativos.

`auto`A política é necessária para impulsionar a disposição em camadas desses blocos, porque somente a `auto` política afeta os blocos que estão no sistema de arquivos ativo.

Qualquer tipo de acesso aos dados repõe os dados do mapa de calor. A verificação de vírus, a indexação e até mesmo a atividade de backup que lê os arquivos de origem impedem a categorização porque o limite necessário `tiering-minimum-cooling-days` nunca é atingido.

Disposição em camadas de arquivos parcial

Como o FabricPool funciona no nível de bloco, os arquivos sujeitos a alteração podem ser parcialmente dispostos em camadas no storage de objetos, permanecendo também parcialmente na camada de performance.

Isso é comum com bancos de dados. Bancos de dados que contêm blocos inativos também são candidatos à disposição em camadas do FabricPool. Por exemplo, um banco de dados de gerenciamento da cadeia de suprimentos pode conter informações históricas que devem estar disponíveis se necessário, mas não são acessadas durante operações normais. O FabricPool pode ser usado para realocar seletivamente os blocos inativos.

Por exemplo, os arquivos de dados executados em um volume FabricPool com `tiering-minimum-cooling-days` um período de 90 dias mantêm todos os blocos acessados nos 90 dias anteriores no nível de performance. No entanto, qualquer coisa que não seja acessada por 90 dias é realocada para o nível de capacidade. Em outros casos, a atividade normal da aplicação preserva os blocos corretos no nível correto. Por exemplo, se um banco de dados é normalmente usado para processar os 60 dias anteriores de dados regularmente, um período muito menor `tiering-minimum-cooling-days` pode ser definido porque a atividade natural do aplicativo garante que os blocos não sejam transferidos prematuramente.

A `auto` política deve ser usada com cuidado com bancos de dados. Muitos bancos de dados têm atividades periódicas, como processo de fim de trimestre ou operações de reindexação. Se o período dessas operações for maior do que os `tiering-minimum-cooling-days` problemas de desempenho podem ocorrer. Por exemplo, se o processamento no final do trimestre exigir 1TB TB de dados que, de outra forma, foram intocados, esses dados podem agora estar presentes na camada de capacidade. As leituras do nível de capacidade geralmente são extremamente rápidas e podem não causar problemas de desempenho, mas os resultados exatos dependerão da configuração do armazenamento de objetos.

Políticas

A `tiering-minimum-cooling-days` política deve ser definida suficientemente alta para reter arquivos que possam ser necessários no nível de desempenho. Por exemplo, um banco de dados em que os 60 dias mais recentes de dados possam ser necessários com desempenho ideal garantiria a definição `tiering-minimum-cooling-days` do período para 60 dias. Resultados semelhantes também podem ser alcançados com base nos padrões de acesso dos arquivos. Por exemplo, se os 90 dias mais recentes de dados forem necessários e o aplicativo estiver acessando esse período de 90 dias de dados, os dados permanecerão no nível de desempenho. Definir o `tiering-minimum-cooling-days` período para 2 dias classificaria os dados imediatamente após os dados ficarem menos ativos.

`auto`A política é necessária para impulsionar a disposição em camadas desses blocos, porque somente a `auto` política afeta os blocos que estão no sistema de arquivos ativo.

Qualquer tipo de acesso aos dados repõe os dados do mapa de calor. Portanto, varreduras completas de tabela de banco de dados e até mesmo atividades de backup que leem os arquivos de origem impedem a categorização porque o limite necessário `tiering-minimum-cooling-days` nunca é atingido.

Disposição em camadas do log de arquivamento

Talvez o uso mais importante para o FabricPool seja melhorar a eficiência de dados inativos conhecidos, como logs de transações de banco de dados.

A maioria dos bancos de dados relacionais opera no modo de arquivamento de log de transações para fornecer recuperação pontual. As alterações nos bancos de dados são confirmadas registrando as alterações nos logs de transação e o log de transação é mantido sem ser substituído. O resultado pode ser um requisito para manter um enorme volume de logs de transações arquivados. Exemplos semelhantes existem com muitos outros fluxos de trabalho de aplicações que geram dados que precisam ser retidos, mas é altamente improvável que nunca sejam acessados.

A FabricPool soluciona esses problemas ao fornecer uma única solução com disposição em camadas integrada. Os arquivos são armazenados e permanecem acessíveis em seu local habitual, mas praticamente não ocupam espaço na matriz primária.

Políticas

O uso de uma `tiering-minimum-cooling-days` política de poucos dias resulta na retenção de blocos nos arquivos criados recentemente (que são os arquivos mais prováveis de serem necessários no curto prazo) na categoria de performance. Os blocos de dados de arquivos mais antigos são movidos para o nível de capacidade.

O `auto` aplica a disposição em camadas quando o limite de resfriamento tiver sido atingido, independentemente de os logs terem sido excluídos ou continuarem a existir no sistema de arquivos primário. Armazenar todos os logs potencialmente necessários em um único local no sistema de arquivos ativo também simplifica o gerenciamento. Não há razão para pesquisar instantâneos para localizar um arquivo que precisa ser restaurado.

Alguns aplicativos, como o Microsoft SQL Server, truncam arquivos de log de transações durante operações de backup para que os logs não estejam mais no sistema de arquivos ativo. A capacidade pode ser salva usando a `snapshot-only` política de disposição em camadas, mas a `auto` política não é útil para dados de log, porque raramente devem existir dados de log refrigerados no sistema de arquivos ativo.

Disposição do Snapshot em camadas

A versão inicial do FabricPool visou o caso de uso de backup. O único tipo de blocos que podiam ser dispostos em camadas eram blocos que não estavam mais associados aos dados no sistema de arquivos ativo. Portanto, apenas os blocos de dados do snapshot podem ser movidos para a camada de capacidade. Essa continua sendo uma das opções de disposição em categorias mais seguras quando você precisa garantir que a

performance nunca seja afetada.

Políticas - instantâneos locais

Existem duas opções para a disposição em camadas de blocos snapshot inativos na camada de capacidade. Primeiro, a `snapshot-only` política segmenta apenas os blocos de snapshot. Embora `auto` a política inclua os `snapshot-only` blocos, ela também dispõe blocos do sistema de arquivos ativo. Isso pode não ser desejável.

O `tiering-minimum-cooling-days` valor deve ser definido para um período de tempo que torne os dados que possam ser necessários durante uma restauração disponíveis no nível de desempenho. Por exemplo, a maioria dos cenários de restauração de um banco de dados de produção crítico inclui um ponto de restauração em algum momento nos últimos dias. Definir um `tiering-minimum-cooling-days` valor de 3 garantirá que qualquer restauração do arquivo resulte em um arquivo que imediatamente forneça o máximo de desempenho. Todos os blocos nos arquivos ativos ainda estão presentes no storage rápido sem a necessidade de recuperá-los da camada de capacidade.

Políticas - instantâneos replicados

Um snapshot replicado com o SnapMirror ou o SnapVault que é usado somente para recuperação geralmente deve usar a política FabricPool `all`. Com essa política, os metadados são replicados, mas todos os blocos de dados são imediatamente enviados para a categoria de capacidade, o que proporciona o máximo de performance. A maioria dos processos de recuperação envolve e/S sequenciais, que é inerentemente eficiente. O tempo de recuperação do destino do armazenamento de objetos deve ser avaliado, mas, em uma arquitetura bem projetada, esse processo de recuperação não precisa ser significativamente mais lento do que a recuperação de dados locais.

Se os dados replicados também se destinarem a ser usados para clonagem, a `auto` política é mais apropriada, com um `tiering-minimum-cooling-days` valor que engloba dados que se espera que sejam usados regularmente em um ambiente de clonagem. Por exemplo, o conjunto de trabalho ativo de um banco de dados pode incluir dados lidos ou gravados nos três dias anteriores, mas também pode incluir outros 6 meses de dados históricos. Em caso afirmativo, a `auto` política no destino do SnapMirror torna o conjunto de trabalho disponível no nível de desempenho.

Disposição em camadas do backup

Os backups tradicionais de aplicativos incluem produtos como o Oracle Recovery Manager, que criam backups baseados em arquivos fora do local do banco de dados original.

`tiering-minimum-cooling-days` policy of a few days preserves the most recent backups, and therefore the backups most likely to be required for an urgent recovery situation, on the performance tier. The data blocks of the older files are then moved to the capacity tier.
A `auto` política é a política mais adequada para dados de backup. Isso garante a disposição de camadas de prompt quando o limite de resfriamento tiver sido atingido, independentemente de os arquivos terem sido excluídos ou continuarem existindo no sistema de arquivos primário. Armazenar todos os arquivos potencialmente necessários em um único local no sistema de arquivos ativo também simplifica o gerenciamento. Não há razão para pesquisar instantâneos para localizar um arquivo que precisa ser restaurado.

A `snapshot-only` política poderia ser feita para funcionar, mas essa política só se aplica a blocos que não estão mais no sistema de arquivos ativo. Portanto, os arquivos em um compartilhamento NFS ou SMB devem ser excluídos primeiro antes que os dados possam ser categorizados.

Esta política seria ainda menos eficiente com uma configuração LUN porque a exclusão de um arquivo de um LUN só remove as referências de arquivo dos metadados do sistema de arquivos. Os blocos reais nos LUNs permanecem no lugar até que sejam substituídos. Essa situação pode criar um longo atraso entre o tempo em que um arquivo é excluído e o tempo em que os blocos são substituídos e se tornam candidatos à disposição em camadas. Há algum benefício ao migrar `snapshot-only` os blocos para a categoria de capacidade, mas, no geral, o gerenciamento de dados de backup da FabricPool funciona melhor com essa `auto` política.

> Essa abordagem ajuda os usuários a gerenciar o espaço necessário para backups com mais eficiência, mas o próprio FabricPool não é uma tecnologia de backup. A disposição em camadas dos arquivos de backup no armazenamento de objetos simplifica o gerenciamento porque os arquivos ainda estão visíveis no sistema de storage original, mas os blocos de dados no destino do armazenamento de objetos dependem do sistema de storage original. Se o volume de origem for perdido, os dados do armazenamento de objetos não serão mais utilizáveis.

Interrupções de acesso ao armazenamento de objetos

A disposição em categorias de um conjunto de dados com o FabricPool resulta em uma dependência entre o storage array primário e a categoria de armazenamento de objetos. Há muitas opções de armazenamento de objetos que oferecem níveis variados de disponibilidade. É importante entender o impacto de uma possível perda de conectividade entre o storage array primário e a camada de storage de objetos.

Se uma e/S emitida para o ONTAP exigir dados da camada de capacidade e o ONTAP não puder alcançar o nível de capacidade para recuperar blocos, a e/S eventualmente expirará. O efeito deste tempo limite depende do protocolo utilizado. Em um ambiente NFS, o ONTAP responde com uma resposta EJUKEBOX ou EDELAY, dependendo do protocolo. Alguns sistemas operacionais mais antigos podem interpretar isso como um erro, mas os sistemas operacionais atuais e os níveis de patch atuais do cliente Oracle Direct NFS tratam isso como um erro recuperável e continuam aguardando a conclusão da e/S.

Um tempo limite mais curto se aplica a ambientes SAN. Se um bloco no ambiente de armazenamento de objetos for necessário e permanecer inacessível por dois minutos, um erro de leitura será retornado ao host.

O volume ONTAP e os LUNs permanecem online, mas o sistema operacional do host pode sinalizar o sistema de arquivos como estando em um estado de erro.

A política de problemas de conetividade de storage de objetos `snapshot-only` é menos preocupante porque apenas os dados de backup são dispostos em camadas. Problemas de comunicação atrasariam a recuperação de dados, mas de outra forma não afetariam os dados que estão sendo usados ativamente. As `auto` políticas e `all` permitem a disposição em camadas de dados inativos do LUN ativo, o que significa que um erro durante a recuperação de dados do armazenamento de objetos pode afetar a disponibilidade do banco de dados. Uma implantação de SAN com essas políticas deve ser usada somente com conexões de rede e storage de objetos de classe empresarial projetadas para alta disponibilidade. NetApp StorageGRID é a opção superior.

Proteção de dados Oracle

Proteção de dados com o ONTAP

O NetApp sabe que os dados mais críticos são encontrados em bancos de dados.

Uma empresa não pode operar sem acesso a seus dados e, às vezes, os dados definem o negócio. No entanto, a proteção de dados é mais do que apenas garantir um backup utilizável. No entanto, é preciso realizar os backups de maneira rápida e confiável, além de armazená-los em segurança.

O outro lado da proteção de dados é a recuperação de dados. Quando os dados estão inacessíveis, a empresa é afetada e pode estar inoperante até que os dados sejam restaurados. Este processo deve ser rápido e confiável. Finalmente, a maioria dos bancos de dados deve ser protegida contra desastres, o que significa manter uma réplica do banco de dados. A réplica deve estar suficientemente atualizada. Também deve ser rápido e simples tornar a réplica um banco de dados totalmente operacional.

Esta documentação substitui o relatório técnico publicado anteriormente *TR-4591: Proteção de dados Oracle: Backup, recuperação e replicação.*

Planejamento

A arquitetura de proteção de dados empresariais certa depende dos requisitos de negócios relacionados à retenção de dados, capacidade de recuperação e tolerância de interrupções durante vários eventos.

Por exemplo, considere o número de aplicações, bancos de dados e conjuntos de dados importantes no escopo. Criar uma estratégia de backup para um único conjunto de dados que garanta a conformidade com SLAs típicos é bastante simples, pois não há muitos objetos para gerenciar. À medida que o número de conjuntos de dados aumenta, o monitoramento se torna mais complicado e os administradores podem ser forçados a gastar um tempo cada vez maior lidando com falhas de backup. À medida que um ambiente atinge a nuvem e o provedor de serviços dimensiona, é necessária uma abordagem totalmente diferente.

O tamanho do conjunto de dados também afeta a estratégia. Por exemplo, existem muitas opções para backup e recuperação com um banco de dados 100GB porque o conjunto de dados é tão pequeno. A simples cópia de dados de Mídia de backup com ferramentas tradicionais normalmente fornece um rto suficiente para recuperação. Um banco de dados 100TB normalmente precisa de uma estratégia completamente diferente, a menos que o rto permita uma interrupção de vários dias, caso em que um procedimento tradicional de backup e recuperação baseado em cópia pode ser aceitável.

Finalmente, existem fatores fora do próprio processo de backup e recuperação. Por exemplo, existem bancos de dados que suportam atividades críticas de produção, tornando a recuperação um evento raro que só é realizado por DBAs qualificados? Alternativamente, os bancos de dados fazem parte de um grande ambiente

de desenvolvimento em que a recuperação é uma ocorrência frequente e gerenciada por uma EQUIPE DE TI generalista?

Planejamento de rto, RPO e SLA

O ONTAP permite que você personalize facilmente uma estratégia de proteção de dados de banco de dados Oracle de acordo com suas necessidades de negócios.

Esses requisitos incluem fatores como a velocidade de recuperação, a perda máxima de dados permitida e as necessidades de retenção de backup. O plano de proteção de dados também deve levar em consideração vários requisitos regulatórios para retenção e restauração de dados. Finalmente, diferentes cenários de recuperação de dados devem ser considerados, desde a recuperação típica e previsível resultante de erros de usuário ou aplicativo até cenários de recuperação de desastres que incluem a perda completa de um site.

Pequenas alterações nas políticas de proteção e recuperação de dados podem ter um efeito significativo na arquitetura geral de storage, backup e recuperação. É fundamental definir e documentar padrões antes de iniciar o trabalho de design para evitar complicar uma arquitetura de proteção de dados. Recursos desnecessários ou níveis de proteção levam a custos desnecessários e sobrecarga de gerenciamento, e um requisito inicialmente negligenciado pode levar um projeto na direção errada ou exigir alterações de design de última hora.

Objetivo de tempo de recuperação

O objetivo de tempo de recuperação (rto) define o tempo máximo permitido para a recuperação de um serviço. Por exemplo, um banco de dados de recursos humanos pode ter um rto de 24 horas porque, embora seja muito inconveniente perder o acesso a esses dados durante o dia de trabalho, a empresa ainda pode operar. Em contraste, um banco de dados que suporta o Registro geral de um banco teria um rto medido em minutos ou mesmo segundos. Um rto de zero não é possível, porque deve haver uma maneira de diferenciar entre uma interrupção de serviço real e um evento de rotina, como um pacote de rede perdido. No entanto, um rto quase zero é um requisito típico.

Objetivo do ponto de restauração

O objetivo do ponto de restauração (RPO) define a perda máxima de dados tolerável. Em muitos casos, o RPO é determinado exclusivamente pela frequência de snapshots ou atualizações do SnapMirror.

Em alguns casos, o RPO pode ser tornado mais agressivo, protegendo determinados dados de forma seletiva com mais frequência. Em um contexto de banco de dados, o RPO geralmente é uma questão de quanto dados de log podem ser perdidos em uma situação específica. Em um cenário típico de recuperação em que um banco de dados é danificado devido a um bug de produto ou erro de usuário, o RPO deve ser zero, o que significa que não deve haver perda de dados. O procedimento de recuperação envolve restaurar uma cópia anterior dos arquivos de banco de dados e, em seguida, rereproduzir os arquivos de log para trazer o estado do banco de dados até o ponto desejado no tempo. Os ficheiros de registo necessários para esta operação já devem estar no local original.

Em cenários incomuns, os dados de log podem ser perdidos. Por exemplo, um acidental ou malicioso `rm -rf *` de arquivos de banco de dados pode resultar na exclusão de todos os dados. A única opção seria restaurar do backup, incluindo arquivos de log, e alguns dados seriam inevitavelmente perdidos. A única opção para melhorar o RPO em um ambiente de backup tradicional seria executar backups repetidos dos dados de log. Isso tem limitações, no entanto, devido à constante movimentação de dados e à dificuldade de manter um sistema de backup como um serviço em constante execução. Um dos benefícios dos sistemas avançados de storage é a capacidade de proteger dados de danos acidentais ou mal-intencionados nos arquivos, além de fornecer um RPO melhor sem movimentação de dados.

Recuperação de desastres

A recuperação de desastres inclui a arquitetura, as políticas e os procedimentos DE TI necessários para recuperar um serviço em caso de desastre físico. Isso pode incluir inundações, incêndios ou pessoas agindo com intenção maliciosa ou negligente.

A recuperação de desastres é mais do que apenas um conjunto de procedimentos de recuperação. É o processo completo de identificar os vários riscos, definir os requisitos de recuperação de dados e continuidade do serviço e fornecer a arquitetura certa com os procedimentos associados.

Ao estabelecer os requisitos de proteção de dados, é essencial diferenciar entre os requisitos típicos de RPO e rto e os requisitos de RPO e rto necessários para a recuperação de desastres. Alguns ambientes de aplicações exigem um RPO de zero e um rto quase zero para situações de perda de dados que vão desde um erro de usuário relativamente normal até um incêndio que destrói um data center. No entanto, existem consequências administrativas e de custos para estes elevados níveis de proteçãoão.

Em geral, os requisitos de recuperação de dados que não sejam de desastres devem ser rigorosos por dois motivos. Primeiro, erros de aplicativos e erros de usuário que danificam dados são previsíveis ao ponto de serem quase inevitáveis. Em segundo lugar, não é difícil projetar uma estratégia de backup que possa fornecer um RPO de zero e um rto baixo, contanto que o sistema de storage não seja destruído. Não há motivo para não solucionar um risco significativo que seja facilmente remediado, e é por isso que os destinos RPO e rto para recuperação local devem ser agressivos.

Os requisitos de rto e RPO para recuperação de desastres variam mais amplamente com base na probabilidade de um desastre e nas consequências da perda ou interrupção de dados associada a uma empresa. Os requisitos de RPO e rto devem ser baseados nas necessidades reais dos negócios e não em princípios gerais. Eles devem levar em conta vários cenários de desastre lógico e físico.

Desastres lógicos

Os desastres lógicos incluem corrupção de dados causada por usuários, erros de aplicativos ou sistemas operacionais e falhas de software. Os desastres lógicos também podem incluir ataques maliciosos de terceiros com vírus ou worms ou explorando vulnerabilidades de aplicativos. Nesses casos, a infraestrutura física não está danificada, mas os dados subjacentes não são mais válidos.

Um tipo cada vez mais comum de desastre lógico é conhecido como ransomware, no qual um vetor de ataque é usado para criptografar dados. A criptografia não danifica os dados, mas torna-os indisponíveis até que o pagamento seja feito a terceiros. Um número cada vez maior de empresas está sendo alvo específico de hacks de ransomware. Para essa ameaça, o NetApp oferece snapshots à prova de violações, onde nem mesmo o administrador de storage pode alterar os dados protegidos antes da data de expiração configurada.

Desastres físicos

Os desastres físicos incluem a falha de componentes de uma infraestrutura que excede seus recursos de redundância e resultam em perda de dados ou perda estendida de serviço. Por exemplo, a proteção RAID fornece redundância de unidade de disco e o uso de HBAs fornece redundância de porta FC e cabo FC. Falhas de hardware de tais componentes são previsíveis e não afetam a disponibilidade.

Em um ambiente corporativo, geralmente é possível proteger a infraestrutura de um local inteiro com componentes redundantes até o ponto em que o único cenário de desastre físico previsível é a perda completa do local. O Planejamento de recuperação de desastre depende da replicação local a local.

Proteção de dados síncrona e assíncrona

Em um mundo ideal, todos os dados seriam replicados de forma síncrona em locais geograficamente

dispersos. Essa replicação nem sempre é viável ou até possível por várias razões:

- A replicação síncrona aumenta inevitavelmente a latência de gravação porque todas as alterações devem ser replicadas em ambos os locais antes que a aplicação/banco de dados possa prosseguir com o processamento. O efeito de desempenho resultante às vezes é inaceitável, descartando o uso do espelhamento síncrono.
- O aumento da adoção do storage SSD de 100% significa que a latência de gravação adicional provavelmente será notada porque as expectativas de desempenho incluem centenas de milhares de IOPS e latência inferior a milissegundos. Aproveitar todos os benefícios do uso de SSDs de 100% pode exigir uma nova visita à estratégia de recuperação de desastres.
- Os conjuntos de dados continuam a crescer em termos de bytes, criando desafios com a garantia de largura de banda suficiente para sustentar a replicação síncrona.
- Os conjuntos de dados também crescem em termos de complexidade, criando desafios com o gerenciamento da replicação síncrona de grande escala.
- Estratégias baseadas na nuvem frequentemente envolvem maiores distâncias de replicação e latência, o que impede o uso do espelhamento síncrono.

A NetApp oferece soluções que incluem replicação síncrona para as demandas mais exigentes de recuperação de dados e soluções assíncronas que proporcionam melhor desempenho e flexibilidade. Além disso, a tecnologia NetApp se integra perfeitamente a muitas soluções de replicação de terceiros, como o Oracle DataGuard

Tempo de retenção

O último aspeto de uma estratégia de proteção de dados é o tempo de retenção de dados, que pode variar muito.

- Um requisito típico é de 14 dias de backups noturnos no local principal e 90 dias de backups armazenados em um local secundário.
- Muitos clientes criam arquivos trimestrais autônomos armazenados em Mídias diferentes.
- Um banco de dados constantemente atualizado pode não ter necessidade de dados históricos, e os backups precisam ser mantidos apenas por alguns dias.
- Os requisitos regulatórios podem exigir recuperação até o ponto de qualquer transação arbitrária em uma janela de 365 dias.

Disponibilidade do banco de dados

O ONTAP foi projetado para oferecer a máxima disponibilidade de banco de dados Oracle. Uma descrição completa dos recursos de alta disponibilidade do ONTAP está além do escopo deste documento. No entanto, assim como na proteção de dados, uma compreensão básica dessa funcionalidade é importante ao projetar uma infraestrutura de banco de dados.

Pares HA

A unidade básica de alta disponibilidade é o par de HA. Cada par contém links redundantes para suportar a replicação de dados para o NVRAM. NVRAM não é um cache de gravação. A RAM dentro do controlador serve como cache de gravação. O objetivo do NVRAM é registrar temporariamente os dados como uma salvaguarda contra falhas inesperadas do sistema. A este respeito, é semelhante a um log refazer banco de dados.

Tanto o NVRAM quanto o log refazer de banco de dados são usados para armazenar dados rapidamente, permitindo que as alterações aos dados sejam confirmadas o mais rápido possível. A atualização para os dados persistentes em unidades (ou datafiles) não ocorre até mais tarde durante um processo chamado ponto de verificação em plataformas ONTAP e na maioria dos bancos de dados. Nem os dados do NVRAM nem os logs de refazer do banco de dados são lidos durante operações normais.

Se um controlador falhar abruptamente, é provável que haja alterações pendentes armazenadas no NVRAM que ainda não foram gravadas nas unidades. O controlador do parceiro deteta a falha, controla as unidades e aplica as alterações necessárias que foram armazenadas no NVRAM.

Takeover e giveback

Takeover e giveback refere-se ao processo de transferência de responsabilidade por recursos de storage entre nós em um par de HA. Há dois aspetos para a aquisição e a giveback:

- Gestão da conetividade de rede que permite o acesso às unidades
- Gestão das próprias unidades

As interfaces de rede que suportam o tráfego CIFS e NFS são configuradas com um local de origem e failover. Uma aquisição inclui mover as interfaces de rede para sua casa temporária em uma interface física localizada na(s) mesma(s) sub-rede(s) do local original. A giveback inclui mover as interfaces de rede de volta para seus locais originais. O comportamento exato pode ser ajustado conforme necessário.

As interfaces de rede que suportam protocolos de bloco SAN, como iSCSI e FC, não são relocadas durante a aquisição e a giveback. Em vez disso, os LUNs devem ser provisionados com caminhos que incluam um par de HA completo, o que resulta em um caminho primário e um caminho secundário.

Caminhos adicionais para controladores adicionais também podem ser configurados para dar suporte à realocação de dados entre nós em um cluster maior, mas isso não faz parte do processo de HA.

O segundo aspeto da aquisição e da giveback é a transferência da propriedade do disco. O processo exato depende de vários fatores, incluindo o motivo da aquisição/giveback e as opções de linha de comando emitidas. O objetivo é realizar a operação da forma mais eficiente possível. Embora o processo geral possa parecer exigir vários minutos, o momento real em que a propriedade da unidade é transferida de nó para nó geralmente pode ser medido em segundos.

Tempo de takeover

A e/S do host tem uma pequena pausa na e/S durante as operações de takeover e giveback, mas não deve haver interrupção do aplicativo em um ambiente configurado corretamente. O processo de transição real no qual a e/S é atrasada geralmente é medido em segundos, mas o host pode exigir tempo adicional para reconhecer a alteração nos caminhos de dados e reenviar operações de e/S.

A natureza da interrupção depende do protocolo:

- Uma interface de rede que suporta tráfego NFS e CIFS emite uma solicitação ARP (Address Resolution Protocol) para a rede após a transição para um novo local físico. Isso faz com que os switches de rede atualizem suas tabelas de endereços de controle de acesso de Mídia (MAC) e retomem a e/S de processamento A interrupção no caso de aquisição planejada e a giveback geralmente é medida em segundos e, em muitos casos, não é detetável. Algumas redes podem ser mais lentas para reconhecer completamente a mudança no caminho da rede, e alguns sistemas operacionais podem colocar em fila muita e/S em um tempo muito curto que deve ser tentado novamente. Isto pode prolongar o tempo necessário para retomar a I/O.

- Uma interface de rede que suporte protocolos SAN não faz a transição para um novo local. Um sistema operacional do host deve alterar o caminho ou os caminhos em uso. A pausa em I/o observada pelo host depende de vários fatores. Do ponto de vista do sistema de armazenamento, o período em que a e/S não pode ser atendida é de apenas alguns segundos. No entanto, diferentes sistemas operacionais de host podem exigir tempo adicional para permitir que uma e/S termine o tempo limite antes de tentar novamente. Os sistemas operacionais mais novos são mais capazes de reconhecer uma mudança de caminho muito mais rapidamente, mas os sistemas operacionais mais antigos geralmente exigem até 30 segundos para reconhecer uma mudança.

Os tempos de aquisição esperados durante os quais o sistema de storage não pode fornecer dados a um ambiente de aplicativo são mostrados na tabela abaixo. Não deve haver erros em qualquer ambiente de aplicativo, o controle deve aparecer como uma pequena pausa no processamento de e/S.

	NFS	AFF	ASA
Takeover planejado	15 seg	6-10 seg	2-3 seg
Takeover não planejado	30 seg	6-10 seg	2-3 seg

Somas de verificação e integridade de dados

O ONTAP e seus protocolos compatíveis incluem vários recursos que protegem a integridade do banco de dados Oracle, incluindo dados em repouso e dados transmitidos pela rede de rede.

A proteção de dados lógicos no ONTAP consiste em três requisitos principais:

- Os dados devem estar protegidos contra corrupção de dados.
- Os dados devem estar protegidos contra falha da unidade.
- As alterações nos dados devem ser protegidas contra perda.

Essas três necessidades são discutidas nas seções a seguir.

Corrupção de rede: Somas de verificação

O nível mais básico de proteção de dados é a soma de verificação, que é um código especial de deteção de erros armazenado junto com os dados. A corrupção de dados durante a transmissão da rede é detetada com o uso de uma soma de verificação e, em alguns casos, várias somas de verificação.

Por exemplo, um quadro FC inclui uma forma de checksum chamada de verificação de redundância cíclica (CRC) para garantir que a carga útil não esteja corrompida em trânsito. O transmissor envia os dados e o CRC dos dados. O recetor de um quadro FC recalcula o CRC dos dados recebidos para garantir que ele corresponda ao CRC transmitido. Se o CRC recém-calculado não corresponder ao CRC anexado ao quadro, os dados estarão corrompidos e o quadro FC será descartado ou rejeitado. Uma operação de e/S iSCSI inclui somas de verificação nas camadas TCP/IP e Ethernet e, para proteção adicional, também pode incluir proteção CRC opcional na camada SCSI. Qualquer corrupção de bits no fio é detetada pela camada TCP ou camada IP, o que resulta na retransmissão do pacote. Assim como no FC, erros no CRC SCSI resultam em uma rejeição ou rejeição da operação.

Drive corruption: Somas de verificação

As somas de verificação também são usadas para verificar a integridade dos dados armazenados nas unidades. Os blocos de dados gravados nas unidades são armazenados com uma função de checksum que

produz um número imprevisível vinculado aos dados originais. Quando os dados são lidos a partir da unidade, a soma de verificação é recalculada e comparada com a soma de verificação armazenada. Se não corresponder, os dados ficaram corrompidos e devem ser recuperados pela camada RAID.

Corrupção de dados: Gravações perdidas

Um dos tipos mais difíceis de detetar é uma gravação perdida ou extraviada. Quando uma escrita é reconhecida, ela deve ser escrita para a Mídia no local correto. A corrupção de dados no local é relativamente fácil de detetar usando uma soma de verificação simples armazenada com os dados. No entanto, se a gravação é simplesmente perdida, então a versão anterior dos dados ainda pode existir e a soma de verificação estaria correta. Se a gravação for colocada no local físico errado, a soma de verificação associada seria mais uma vez válida para os dados armazenados, mesmo que a gravação tenha destruído outros dados.

A solução para este desafio é a seguinte:

- Uma operação de gravação deve incluir metadados que indicam o local onde a gravação deve ser encontrada.
- Uma operação de gravação deve incluir algum tipo de identificador de versão.

Quando o ONTAP grava um bloco, ele inclui dados sobre a localização do bloco. Se uma leitura subsequente identificar um bloco, mas os metadados indicarem que ele pertence ao local 123 quando foi encontrado no local 456, então a gravação foi extraviada.

Detetar uma escrita totalmente perdida é mais difícil. A explicação é muito complicada, mas essencialmente o ONTAP está armazenando metadados de uma forma que uma operação de gravação resulta em atualizações para dois locais diferentes nas unidades. Se uma gravação for perdida, uma leitura posterior dos dados e metadados associados mostrará duas identidades de versão diferentes. Isso indica que a gravação não foi concluída pela unidade.

A corrupção de gravação perdida e extraviada é extremamente rara, mas, à medida que as unidades continuam a crescer e os conjuntos de dados aumentam a escala de exabytes, o risco aumenta. A detecção de gravações perdidas deve ser incluída em qualquer sistema de storage que suporte workloads de banco de dados.

Falhas de unidade: RAID, RAID DP e RAID-teC

Se um bloco de dados em uma unidade for descoberto como corrompido ou toda a unidade falhar e estiver totalmente indisponível, os dados devem ser reconstituídos. Isso é feito no ONTAP usando unidades de paridade. Os dados são distribuídos em várias unidades de dados e, em seguida, os dados de paridade são gerados. Este é armazenado separadamente dos dados originais.

O ONTAP usou originalmente o RAID 4, que usa uma única unidade de paridade para cada grupo de unidades de dados. O resultado foi que qualquer unidade do grupo poderia falhar sem resultar em perda de dados. Se a unidade de paridade falhar, nenhum dado será danificado e uma nova unidade de paridade poderá ser construída. Se uma única unidade de dados falhar, as unidades restantes poderão ser usadas com a unidade de paridade para regenerar os dados em falta.

Quando as unidades eram pequenas, a chance estatística de duas unidades falharem simultaneamente foi insignificante. À medida que as capacidades da unidade crescem, também tem o tempo necessário para reconstruir os dados após uma falha de unidade. Isso aumentou a janela na qual uma segunda falha da unidade resultaria em perda de dados. Além disso, o processo de reconstrução cria muitas e/S adicionais nas unidades sobreviventes. À medida que as unidades envelhecem, o risco de carga adicional que leva a uma segunda falha da unidade também aumenta. Finalmente, mesmo que o risco de perda de dados não tenha aumentado com o uso continuado do RAID 4, as consequências da perda de dados se tornariam mais graves. Quanto mais dados forem perdidos no caso de uma falha do grupo RAID, mais tempo levaria para recuperar

os dados, estendendo a interrupção dos negócios.

Esses problemas levaram a NetApp a desenvolver a tecnologia NetApp RAID DP, uma variante do RAID 6. Essa solução inclui duas unidades de paridade, o que significa que todas as duas unidades em um grupo RAID podem falhar sem criar perda de dados. As unidades continuaram a crescer em tamanho, o que levou a NetApp a desenvolver a tecnologia NetApp RAID-teC, que introduz uma terceira unidade de paridade.

Algumas práticas recomendadas de banco de dados históricos recomendam o uso do RAID-10, também conhecido como espelhamento distribuído. Isso oferece menos proteção de dados do que até mesmo o RAID DP porque há vários cenários de falha de dois discos, enquanto que no RAID DP não há nenhum.

Há também algumas práticas recomendadas de banco de dados históricos que indicam que RAID-10 é preferível às opções RAID-4/5/6 devido a problemas de desempenho. Essas recomendações às vezes se referem a uma penalidade de RAID. Embora essas recomendações geralmente estejam corretas, elas não são aplicáveis às implementações de RAID no ONTAP. O problema de desempenho está relacionado com a regeneração de paridade. Com as implementações tradicionais de RAID, o processamento das gravações aleatórias de rotina executadas por um banco de dados requer várias leituras de disco para regenerar os dados de paridade e concluir a gravação. A penalidade é definida como o IOPS de leitura adicional necessário para executar operações de gravação.

O ONTAP não incorre em uma penalidade de RAID porque as gravações são encenadas na memória em que a paridade é gerada e, em seguida, gravadas no disco como um único stripe RAID. Não são necessárias leituras para concluir a operação de gravação.

Em resumo, quando comparado ao RAID 10, o RAID DP e o RAID-teC oferecem muito mais capacidade utilizável, melhor proteção contra falha de unidade e nenhum sacrifício de performance.

Proteção contra falhas de hardware: NVRAM

Qualquer storage array que atenda a um workload de banco de dados precisa atender às operações de gravação o mais rápido possível. Além disso, uma operação de gravação deve ser protegida contra perda de um evento inesperado, como uma falha de energia. Isso significa que qualquer operação de gravação deve ser armazenada com segurança em pelo menos dois locais.

Os sistemas AFF e FAS contam com a NVRAM para atender a esses requisitos. O processo de escrita funciona da seguinte forma:

1. Os dados de gravação de entrada são armazenados na RAM.
2. As alterações que devem ser feitas nos dados no disco são registradas no NVRAM no nó local e no nó parceiro. O NVRAM não é um cache de gravação; em vez disso, é um diário semelhante a um log de refazer de banco de dados. Em condições normais, não é lido. Ele é usado apenas para recuperação, como após uma falha de energia durante o processamento de e/S.
3. A gravação é então reconhecida para o host.

O processo de gravação nesta fase é concluído do ponto de vista da aplicação, e os dados são protegidos contra perda porque são armazenados em dois locais diferentes. Eventualmente, as alterações são gravadas no disco, mas esse processo está fora da banda do ponto de vista do aplicativo, porque ocorre depois que a gravação é reconhecida e, portanto, não afeta a latência. Este processo é mais uma vez semelhante ao log de banco de dados. Uma alteração ao banco de dados é registrada nos logs de refazer o mais rápido possível, e a alteração é então reconhecida como comprometida. As atualizações para os datafiles ocorrem muito mais tarde e não afetam diretamente a velocidade de processamento.

No caso de uma falha do controlador, o controlador do parceiro assume a propriedade dos discos necessários e replica os dados registados no NVRAM para recuperar quaisquer operações de e/S que estivessem em

trânsito quando a falha ocorreu.

Proteção contra falhas de hardware: 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 o parceiro de HA falhar, esses dados em trânsito 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 de HA. Nenhum dado é perdido porque o controlador que sofre a falha não reconheceu a operação de gravação.

O ONTAP não permite um failover quando os dados estão fora de sincronia, a menos que o failover 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 for forçado porque os bancos de dados mantêm grandes caches internos de dados no disco. Se ocorrer um failover forçado, 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.

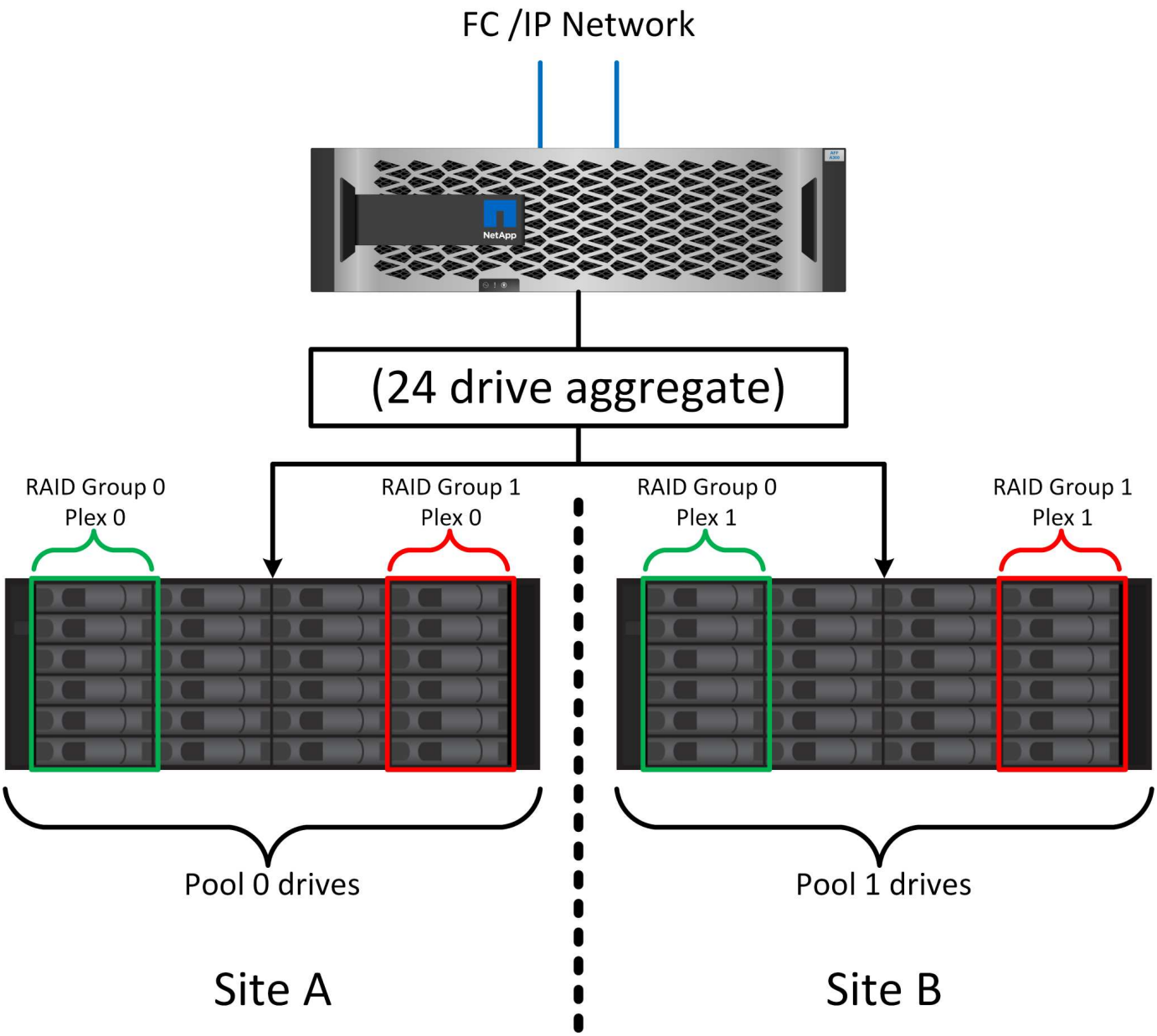
Para proteger os dados contra essa situação, o ONTAP permite que os volumes sejam configurados para proteção especial contra falhas do NVRAM. Quando acionado, esse mecanismo de proteção resulta em um volume entrando em um estado chamado NVFAIL. Esse estado resulta em erros de e/S que causam o desligamento de um aplicativo para que eles não usem dados obsoletos. Os dados não devem ser perdidos porque qualquer gravação reconhecida deve estar presente no storage array.

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.

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:

- 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..
- 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..
- 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. Isso fornece proteção contínua 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 Group 0 inclui um Plex de 6 unidades no local Um espelhado para um Plex de 6 unidades no local B. da mesma forma, RAID Group 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 por alimentações de energia independentes de sistemas UPS independentes.

Somas de verificação

O tópico de checksums é de particular interesse para DBAs que estão acostumados a usar backups de streaming Oracle RMAN migram para backups baseados em snapshot. Um recurso do RMAN é que ele executa verificações de integridade durante operações de backup. Embora esse recurso tenha algum valor, seu principal benefício é para um banco de dados que não é usado em um storage array moderno. Quando as unidades físicas são usadas para um banco de dados Oracle, é quase certo que a corrupção eventualmente ocorre à medida que as unidades envelhecem, um problema que é resolvido por somas de verificação baseadas em array em arrays de armazenamento reais.

Com um storage array real, a integridade de dados é protegida pelo uso de somas de verificação em vários níveis. Se os dados estiverem corrompidos em uma rede baseada em IP, a camada TCP (Transmission Control Protocol) rejeita os dados do pacote e solicita a retransmissão. O protocolo FC inclui somas de verificação, assim como os dados SCSI encapsulados. Depois que ele está no array, o ONTAP tem proteção RAID e checksum. A corrupção pode ocorrer, mas, como na maioria dos arrays corporativos, ela é detetada e corrigida. Normalmente, uma unidade inteira falha, solicitando uma reconstrução RAID e a integridade do banco de dados não é afetada. Ainda é possível que bytes individuais em uma unidade sejam danificados por radiação cósmica ou células flash com falha. Se isso acontecer, a verificação de paridade falharia, a unidade falharia e uma reconstrução RAID começaria. Mais uma vez, a integridade dos dados não é afetada. A linha final de defesa é o uso de somas de verificação. Se, por exemplo, um erro de firmware catastrófico em uma unidade corrompeu dados de uma forma que de alguma forma não fosse detetado por uma verificação de paridade RAID, a soma de verificação não corresponderia e o ONTAP impediria a transferência de um bloco corrompido antes que o banco de dados Oracle pudesse recebê-lo.

A arquitetura Oracle datafile e refazer log também foi projetada para fornecer o mais alto nível possível de integridade de dados, mesmo em circunstâncias extremas. No nível mais básico, os blocos Oracle incluem checksum e verificações lógicas básicas com quase todas as I/O. Se o Oracle não travou ou uma tablespace off-line, então os dados estão intactos. O grau de verificação da integridade dos dados é ajustável, e o Oracle também pode ser configurado para confirmar gravações. Como resultado, quase todos os cenários de falha e falha podem ser recuperados e, no caso extremamente raro de uma situação irrecuperável, a corrupção é imediatamente detetada.

A maioria dos clientes NetApp que usam bancos de dados Oracle descontinuam o uso de RMAN e outros produtos de backup após a migração para backups baseados em snapshot. Ainda existem opções nas quais o RMAN pode ser usado para executar a recuperação em nível de bloco com o SnapCenter. No entanto, no dia a dia, o RMAN, o NetBackup e outros produtos só são usados ocasionalmente para criar cópias de arquivamento mensais ou trimestrais.

Alguns clientes optam por executar `dbv` periodicamente para realizar verificações de integridade em seus bancos de dados existentes. NetApp desencoraja essa prática porque cria carga de e/S desnecessária. Como discutido acima, se o banco de dados não estava enfrentando problemas anteriormente, a chance de `dbv` detetar um problema é perto de zero, e este utilitário cria uma carga de e/S sequencial muito alta na rede e no sistema de armazenamento. A menos que haja razão para acreditar que existe corrupção, como a exposição a um bug conhecido da Oracle, não há razão para ser executado `dbv`.

Noções básicas de backup e recuperação

Backups baseados em snapshot

A base da proteção de dados de banco de dados Oracle no ONTAP é a tecnologia NetApp Snapshot.

Os valores-chave são os seguintes:

- **Simplicidade.** Um snapshot é uma cópia somente leitura do conteúdo de um contentor de dados em um determinado momento.
- **Eficiência.** Os instantâneos não requerem espaço no momento da criação. O espaço só é consumido quando os dados são alterados.
- **Capacidade de gerenciamento.** Uma estratégia de backup baseada em snapshots é fácil de configurar e gerenciar, pois os snapshots são uma parte nativa do sistema operacional de storage. Se o sistema de armazenamento estiver ligado, ele estará pronto para criar backups.
- **Escalabilidade.** É possível preservar até 1024 backups de um único contêiner de arquivos e LUNs. Para conjuntos de dados complexos, vários contêineres de dados podem ser protegidos por um único conjunto consistente de snapshots.
- O desempenho não é afetado, independentemente de um volume conter 1024 snapshots ou nenhum.

Embora muitos fornecedores de storage ofereçam tecnologia Snapshot, a tecnologia Snapshot no ONTAP é única e oferece benefícios significativos para ambientes de aplicações e bancos de dados empresariais:

- As cópias snapshot fazem parte do layout de arquivo em qualquer lugar (WAFL) subjacente. Eles não são uma tecnologia adicional ou externa. Isso simplifica o gerenciamento porque o sistema de storage é o sistema de backup.
- As cópias snapshot não afetam a performance, exceto em alguns casos de borda, como quando muitos dados são armazenados em snapshots que o sistema de storage subjacente preenche.
- O termo "grupo de consistência" é frequentemente usado para se referir a um agrupamento de objetos de armazenamento que são gerenciados como uma coleta consistente de dados. Um snapshot de um determinado volume ONTAP constitui um backup de grupo de consistência.

Os snapshots do ONTAP também são mais dimensionados do que a tecnologia da concorrência. Os clientes podem armazenar snapshots 5, 50 ou 500 sem afetar a performance. O número máximo de instantâneos atualmente permitido em um volume é 1024. Se a retenção adicional de snapshot for necessária, há opções para colocar os snapshots em cascata em volumes adicionais.

Como resultado, proteger um conjunto de dados hospedado no ONTAP é simples e altamente dimensionável. Os backups não exigem movimentação de dados, portanto, uma estratégia de backup pode ser adaptada às necessidades da empresa, em vez das limitações de taxas de transferência de rede, grande número de unidades de fita ou áreas de preparo de disco.

Um instantâneo é um backup?

Uma pergunta comumente feita sobre o uso de snapshots como estratégia de proteção de dados é o fato de que os dados "reais" e os dados instantâneos estão localizados nas mesmas unidades. A perda dessas unidades resultaria na perda dos dados primários e do backup.

Este é um problema válido. Os snapshots locais são usados para necessidades diárias de backup e recuperação e, nesse aspeto, o snapshot é um backup. Cerca de 99% de todos os cenários de recuperação em ambientes NetApp dependem de snapshots para atender até aos requisitos mais agressivos de rto.

No entanto, os snapshots locais nunca devem ser a única estratégia de backup. É por isso que a NetApp oferece tecnologia como replicação SnapMirror e SnapVault para replicar snapshots para um conjunto independente de unidades com rapidez e eficiência. Em uma solução de arquitetura adequada com snapshots e replicação de snapshot, o uso da fita pode ser minimizado para talvez um arquivamento trimestral ou eliminado totalmente.

Backups baseados em snapshot

Há muitas opções para o uso de cópias ONTAP Snapshot para proteger seus dados. Além disso, os snapshots são a base de muitos outros recursos da ONTAP, incluindo replicação, recuperação de desastres e clonagem. Uma descrição completa da tecnologia de instantâneos está além do escopo deste documento, mas as seções a seguir fornecem uma visão geral.

Existem duas abordagens principais para criar um instantâneo de um conjunto de dados:

- Backups consistentes com falhas
- Backups consistentes com aplicativos

Um backup consistente com falhas de um conjunto de dados refere-se à captura de toda a estrutura do conjunto de dados em um único ponto no tempo. Se o conjunto de dados for armazenado em um único volume, o processo será simples. É possível criar um Snapshot a qualquer momento. Se um conjunto de dados abranger volumes, é necessário criar um instantâneo de grupo de consistência (CG). Existem várias opções para criar snapshots CG, incluindo o software NetApp SnapCenter, recursos nativos do grupo de consistência do ONTAP e scripts mantidos pelo usuário.

Backups consistentes com falhas são usados principalmente quando a recuperação do ponto de backup é suficiente. Quando uma recuperação mais granular é necessária, backups consistentes com aplicações geralmente são necessários.

A palavra "consistente" em "consistente com aplicativos" é muitas vezes um erro. Por exemplo, colocar um banco de dados Oracle no modo de backup é referido como um backup consistente com aplicativos, mas os dados não são consistentes ou silenciosos de forma alguma. Os dados continuam a mudar durante todo o backup. Em contraste, a maioria dos backups MySQL e Microsoft SQL Server realmente silenciam os dados antes de executar o backup. A VMware pode ou não tornar certos arquivos consistentes.

Grupos de consistência

O termo "grupo de consistência" refere-se à capacidade de um storage array gerenciar vários recursos de armazenamento como uma única imagem. Por exemplo, um banco de dados pode consistir em 10 LUNs. O array deve ser capaz de fazer backup, restaurar e replicar esses 10 LUNs de maneira consistente. A restauração não é possível se as imagens dos LUNs não forem consistentes no ponto de backup. A replicação desses 10 LUNs requer que todas as réplicas sejam perfeitamente sincronizadas umas com as outras.

O termo "grupo de consistência" não é frequentemente usado quando se discute ONTAP porque consistência sempre foi uma função básica da arquitetura de volume e agregado dentro do ONTAP. Muitos outros storage arrays gerenciam LUNs ou sistemas de arquivos como unidades individuais. Eles poderiam, então, ser opcionalmente configurados como um "grupo de consistência" para fins de proteção de dados, mas esta é uma etapa extra na configuração.

O ONTAP sempre foi capaz de capturar imagens de dados locais e replicados consistentes. Embora os vários volumes em um sistema ONTAP não sejam geralmente formalmente descritos como um grupo de consistência, é isso que eles são. Um snapshot desse volume é uma imagem de grupo de consistência, a restauração para esse snapshot é uma restauração de grupo de consistência, e o SnapMirror e o SnapVault oferecem replicação de grupo de consistência.

Instantâneos do grupo de consistência

Os snapshots de grupos de consistência (snapshots cg) são uma extensão da tecnologia básica do ONTAP Snapshot. Uma operação de snapshot padrão cria uma imagem consistente de todos os dados em um único volume, mas às vezes é necessário criar um conjunto consistente de snapshots em vários volumes e até

mesmo em vários sistemas de storage. O resultado é um conjunto de instantâneos que podem ser usados da mesma forma que um instantâneo de apenas um volume individual. Eles podem ser usados para recuperação de dados local, replicados para fins de recuperação de desastres ou clonados como uma única unidade consistente.

O maior uso conhecido de snapshots cg é para um ambiente de banco de dados de aproximadamente 1PB TB de tamanho abrangendo 12 controladoras. Os snapshots cg criados neste sistema foram usados para backup, recuperação e clonagem.

Na maioria das vezes, quando um conjunto de dados abrange volumes e ordem de gravação deve ser preservado, um cg-snapshot é usado automaticamente pelo software de gerenciamento escolhido. Nesses casos, não há necessidade de entender os detalhes técnicos dos instantâneos cg. No entanto, há situações em que requisitos complicados de proteção de dados exigem controle detalhado sobre o processo de replicação e proteção de dados. Fluxos de trabalho de automação ou o uso de scripts personalizados para chamar APIs cg-snapshot são algumas das opções. Compreender a melhor opção e o papel do cg-snapshot requer uma explicação mais detalhada da tecnologia.

A criação de um conjunto de instantâneos cg é um processo de duas etapas:

1. Estabeleça cercas de gravação em todos os volumes de destino.
2. Crie instantâneos desses volumes enquanto estiver no estado cercado.

A esgrima de escrita é estabelecida em série. Isso significa que, à medida que o processo de esgrima é configurado em vários volumes, a e/S de gravação é congelada no primeiro volume da sequência, uma vez que continua a ser comprometida com volumes que aparecem mais tarde. Isso pode inicialmente parecer violar o requisito para que a ordem de gravação seja preservada, mas isso só se aplica a e/S que é emitida assincronamente no host e não depende de outras gravações.

Por exemplo, um banco de dados pode emitir muitas atualizações assíncronas de arquivos de dados e permitir que o sistema operacional reordene a e/S e as complete de acordo com sua própria configuração de agendador. A ordem deste tipo de e/S não pode ser garantida porque a aplicação e o sistema operativo já lançaram a exigência de preservar a ordem de escrita.

Como um exemplo de contador, a maioria das atividades de Registro de banco de dados é síncrona. O banco de dados não prossegue com outras gravações de log até que a e/S seja reconhecida, e a ordem dessas gravações deve ser preservada. Se uma e/S de log chegar em um volume cercado, ela não será reconhecida e a aplicação será bloqueada em outras gravações. Da mesma forma, a e/S de metadados do sistema de arquivos geralmente é síncrona. Por exemplo, uma operação de exclusão de arquivos não deve ser perdida. Se um sistema operacional com um sistema de arquivos xfs excluísse um arquivo e a e/S que atualizasse os metadados do sistema de arquivos xfs para remover a referência a esse arquivo aterrado em um volume cercado, a atividade do sistema de arquivos pausará. Isso garante a integridade do sistema de arquivos durante operações cg-snapshot.

Depois que o grima de gravação é configurado nos volumes de destino, eles estão prontos para a criação de snapshot. Os instantâneos não precisam ser criados exatamente ao mesmo tempo porque o estado dos volumes é congelado de um ponto de vista de gravação dependente. Para se proteger contra uma falha na aplicação criando os instantâneos cg, a esgrima de gravação inicial inclui um tempo limite configurável no qual o ONTAP libera automaticamente a esgrima e retoma o processamento de gravação após um número definido de segundos. Se todos os instantâneos forem criados antes do período de tempo limite expirar, o conjunto de instantâneos resultante será um grupo de consistência válido.

Ordem de escrita dependente

Do ponto de vista técnico, a chave para um grupo de consistência é preservar a ordem de gravação e, especificamente, a ordem de gravação dependente. Por exemplo, um banco de dados gravando em 10 LUNs

grava simultaneamente em todos eles. Muitas gravações são emitidas assincronamente, o que significa que a ordem em que são concluídas não é importante e a ordem real que são concluídas varia de acordo com o sistema operacional e o comportamento da rede.

Algumas operações de gravação devem estar presentes no disco antes que o banco de dados possa continuar com gravações adicionais. Essas operações críticas de gravação são chamadas de gravações dependentes. A e/S de gravação subsequente depende da presença dessas gravações no disco. Qualquer snapshot, recuperação ou replicação desses 10 LUNs deve garantir que a ordem de gravação dependente seja garantida. As atualizações do sistema de arquivos são outro exemplo de gravações dependentes da ordem de gravação. A ordem em que as alterações do sistema de arquivos são feitas deve ser preservada ou todo o sistema de arquivos pode ficar corrompido.

Estratégias

Há duas abordagens principais para backups baseados em snapshot:

- Backups consistentes com falhas
- Backups ativos protegidos por snapshot

Um backup consistente com falhas de um banco de dados refere-se à captura de toda a estrutura do banco de dados, incluindo datafiles, logs de refazer e arquivos de controle, em um único ponto no tempo. Se o banco de dados for armazenado em um único volume, o processo será simples; uma captura Instantânea pode ser criada a qualquer momento. Se um banco de dados abranger volumes, um snapshot de grupo de consistência (CG) deve ser criado. Existem várias opções para criar snapshots CG, incluindo o software NetApp SnapCenter, recursos nativos do grupo de consistência do ONTAP e scripts mantidos pelo usuário.

Os backups Snapshot consistentes com falhas são usados principalmente quando a recuperação do ponto de backup é suficiente. Registros de arquivo podem ser aplicados em algumas circunstâncias, mas quando uma recuperação pontual mais granular é necessária, um backup on-line é preferível.

O procedimento básico para um backup on-line baseado em snapshot é o seguinte:

1. Coloque a base de dados `backup` no modo.
2. Crie um instantâneo de todos os volumes que hospedam datafiles.
3. Sair `backup` do modo.
4. Execute o comando `alter system archive log current` para forçar o arquivamento de logs.
5. Crie instantâneos de todos os volumes que hospedam os logs do arquivo.

Este procedimento produz um conjunto de instantâneos contendo datafiles no modo de backup e os logs críticos de arquivo gerados no modo de backup. Estes são os dois requisitos para recuperar um banco de dados. Arquivos como arquivos de controle também devem ser protegidos por conveniência, mas o único requisito absoluto é a proteção para arquivos de dados e logs de arquivo.

Embora clientes diferentes possam ter estratégias muito diferentes, quase todas essas estratégias são baseadas nos mesmos princípios descritos abaixo.

Recuperação baseada em Snapshot

Ao projetar layouts de volume para bancos de dados Oracle, a primeira decisão é se usar a tecnologia NetApp SnapRestore baseada em volume (VBSR).

O SnapRestore baseado em volume permite que um volume seja revertido quase instantaneamente para um ponto anterior no tempo. Como todos os dados no volume são revertidos, o VBSR pode não ser apropriado

para todos os casos de uso. Por exemplo, se um banco de dados inteiro, incluindo datafiles, refazer logs e Registros de arquivamento, for armazenado em um único volume e esse volume for restaurado com VBSR, os dados serão perdidos porque o Registro de arquivo mais recente e os dados de refazer são descartados.

VBSR não é necessário para restaurar. Muitos bancos de dados podem ser restaurados usando o SnapRestore de arquivo único (SFSR) baseado em arquivo ou simplesmente copiando arquivos do snapshot de volta para o sistema de arquivos ativo.

O VBSR é preferido quando um banco de dados é muito grande ou quando ele deve ser recuperado o mais rápido possível, e o uso do VBSR requer isolamento dos arquivos de dados. Em um ambiente NFS, os arquivos de dados de um determinado banco de dados devem ser armazenados em volumes dedicados que não estejam contaminados por qualquer outro tipo de arquivo. Em um ambiente SAN, os arquivos de dados devem ser armazenados em LUNs dedicados em volumes dedicados. Se um gerenciador de volumes for usado (incluindo Oracle Automatic Storage Management [ASM]), o grupo de discos também deve ser dedicado a arquivos de dados.

Isolar datafiles desta maneira permite que eles sejam revertidos para um estado anterior sem danificar outros sistemas de arquivos.

Reserva do Snapshot

Para cada volume com dados Oracle em um ambiente SAN, o `percent-snapshot-space` deve ser definido como zero porque reservar espaço para um snapshot em um ambiente LUN não é útil. Se a reserva fracionária estiver definida como 100, um instantâneo de um volume com LUNs requer espaço livre suficiente no volume, excluindo a reserva instantânea, para absorver 100% de rotatividade de todos os dados. Se a reserva fracionária for definida para um valor mais baixo, uma quantidade correspondente menor de espaço livre será necessária, mas sempre exclui a reserva instantânea. Isso significa que o espaço de reserva do snapshot em um ambiente LUN é desperdiçado.

Em um ambiente NFS, há duas opções:

- Defina o `percent-snapshot-space` com base no consumo de espaço esperado do instantâneo.
- Defina o `percent-snapshot-space` como zero e gerencie o consumo de espaço ativo e instantâneo coletivamente.

Com a primeira opção, `percent-snapshot-space` é definido para um valor diferente de zero, normalmente em torno de 20%. Este espaço é então escondido do usuário. Esse valor não cria, no entanto, um limite de utilização. Se um banco de dados com uma reserva de 20% sofrer 30% de rotatividade, o espaço instantâneo pode crescer além dos limites da reserva de 20% e ocupar espaço não reservado.

O principal benefício de definir uma reserva para um valor como 20% é verificar se algum espaço está sempre disponível para instantâneos. Por exemplo, um volume 1TB com uma reserva de 20% permitiria apenas que um administrador de banco de dados (DBA) armazenasse 800GB TB de dados. Essa configuração garante pelo menos 200GBMB de espaço para consumo de snapshot.

```
`percent-snapshot-space`Quando está definido como zero, todo o espaço no volume está disponível para o usuário final, o que proporciona melhor visibilidade. O DBA deve entender que, se ele ou ela vir um volume de 1TB TB que aproveita snapshots, esse 1TB TB de espaço será compartilhado entre dados ativos e a rotatividade do Snapshot.
```

Não há preferência clara entre a opção um e a opção dois entre os usuários finais.

ONTAP e snapshots de terceiros

O Oracle Doc ID 604683,1 explica os requisitos para suporte a instantâneos de terceiros e as várias opções disponíveis para operações de backup e restauração.

O fornecedor terceirizado deve garantir que os snapshots da empresa estejam em conformidade com os seguintes requisitos:

- Os snapshots devem ser integrados às operações de restauração e recuperação recomendadas pela Oracle.
- Os snapshots devem ser consistentes com falhas de banco de dados no ponto do snapshot.
- A ordenação de gravação é preservada para cada arquivo dentro de um snapshot.

Os produtos de gerenciamento ONTAP e NetApp da Oracle atendem a esses requisitos.

SnapRestore

A restauração rápida de dados no ONTAP a partir de um snapshot é fornecida pela tecnologia NetApp SnapRestore.

Quando um conjunto de dados essencial não está disponível, as operações de negócios essenciais estão inativas. As fitas podem quebrar, e até mesmo as restaurações de backups baseados em disco podem ser lentas para serem transferidas pela rede. O SnapRestore evita esses problemas fornecendo restauração quase instantânea de conjuntos de dados. Mesmo os bancos de dados em escala de petabyte podem ser completamente restaurados com apenas alguns minutos de esforço.

Existem duas formas de SnapRestore - baseado em arquivo/LUN e baseado em volume.

- Arquivos individuais ou LUNs podem ser restaurados em segundos, seja um arquivo 2TB LUN ou 4KB.
- O volume de arquivos ou LUNs pode ser restaurado em segundos, seja 10GB ou 100TB TB de dados.

Um "contentor de arquivos ou LUNs" normalmente se referiria a um FlexVol volume. Por exemplo, você pode ter 10 LUNs que compõem um grupo de discos LVM em um único volume ou um volume pode armazenar os diretórios base NFS de usuários do 1000. Em vez de executar uma operação de restauração para cada arquivo individual ou LUN, você pode restaurar todo o volume como uma única operação. Esse processo também funciona com contêineres com escalabilidade horizontal que incluem vários volumes, como um FlexGroup ou um Grupo de consistência do ONTAP.

O motivo pelo qual o SnapRestore funciona tão rápido e eficientemente é devido à natureza de um snapshot, que é essencialmente uma visualização paralela somente leitura do conteúdo de um volume em um determinado momento. Os blocos ativos são os blocos reais que podem ser alterados, enquanto o snapshot é uma visualização somente leitura no estado dos blocos que constituem os arquivos e LUNs no momento em que o snapshot foi criado.

O ONTAP só permite acesso somente leitura a dados instantâneos, mas os dados podem ser reativados com o SnapRestore. O instantâneo é reativado como uma visualização de leitura e gravação dos dados, retornando os dados ao seu estado anterior. O SnapRestore pode operar no volume ou no nível do arquivo. A tecnologia é essencialmente a mesma com algumas pequenas diferenças de comportamento.

Volume SnapRestore

O SnapRestore baseado em volume retorna todo o volume de dados para um estado anterior. Essa operação não requer movimentação de dados, o que significa que o processo de restauração é essencialmente instantâneo, embora a operação de API ou CLI possa levar alguns segundos para ser processada. Restaurar

1GB TB de dados não é mais complicado ou demorado do que restaurar 1PB TB de dados. Essa funcionalidade é a principal razão pela qual muitos clientes empresariais migram para os sistemas de storage da ONTAP. Ele fornece um rto medido em segundos, até mesmo para os maiores conjuntos de dados.

Uma desvantagem para o SnapRestore baseado em volume é causada pelo fato de que as alterações dentro de um volume são cumulativas ao longo do tempo. Portanto, cada snapshot e os dados de arquivo ativos dependem das alterações que levam a esse ponto. Reverter um volume para um estado anterior significa descartar todas as alterações subsequentes que foram feitas aos dados. O que é menos óbvio, no entanto, é que isso inclui instantâneos criados posteriormente. Isso nem sempre é desejável.

Por exemplo, um SLA de retenção de dados pode especificar 30 dias de backups noturnos. Restaurar um conjunto de dados para um instantâneo criado há cinco dias com o volume SnapRestore descartaria todos os snapshots criados nos cinco dias anteriores, violando o SLA.

Existem várias opções disponíveis para resolver esta limitação:

1. Os dados podem ser copiados de um snapshot anterior, em vez de executar um SnapRestore de todo o volume. Esse método funciona melhor com conjuntos de dados menores.
2. Um snapshot pode ser clonado em vez de restaurado. A limitação a essa abordagem é que o snapshot de origem é uma dependência do clone. Portanto, ele não pode ser excluído a menos que o clone também seja excluído ou seja dividido em um volume independente.
3. Uso de SnapRestore baseado em arquivos.

File SnapRestore (ficheiro)

O SnapRestore baseado em arquivo é um processo de restauração mais granular baseado em snapshot. Em vez de reverter o estado de um volume inteiro, o estado de um arquivo individual ou LUN é revertido. Não é necessário eliminar instantâneos, nem esta operação cria qualquer dependência de um instantâneo anterior. O ficheiro ou LUN fica imediatamente disponível no volume ativo.

Nenhuma movimentação de dados é necessária durante uma restauração do SnapRestore de um arquivo ou LUN. No entanto, algumas atualizações internas de metadados são necessárias para refletir o fato de que os blocos subjacentes em um arquivo ou LUN agora existem em um snapshot e no volume ativo. Não deve haver efeito no desempenho, mas esse processo bloqueia a criação de snapshots até que ele esteja concluído. A taxa de processamento é de aproximadamente 5Gbps (18TBMB/hora) com base no tamanho total dos arquivos restaurados.

Backups online

Dois conjuntos de dados são necessários para proteger e recuperar um banco de dados Oracle no modo de backup. Note que esta não é a única opção de backup Oracle, mas é a mais comum.

- Um instantâneo dos arquivos de dados no modo de backup
- Os logs de arquivo criados enquanto os datafiles estavam no modo de backup

Se a recuperação completa, incluindo todas as transações confirmadas, é necessário um terceiro item:

- Um conjunto de registos de refazer atuais

Existem várias maneiras de impulsionar a recuperação de um backup on-line. Muitos clientes restauram snapshots usando a CLI do ONTAP e, em seguida, usando o Oracle RMAN ou sqlplus para concluir a recuperação. Isso é especialmente comum com grandes ambientes de produção em que a probabilidade e a

frequência de restaurações de banco de dados são extremamente baixas e qualquer procedimento de restauração é Tratado por um DBA qualificado. Para automação completa, soluções como o NetApp SnapCenter incluem um plug-in Oracle com interfaces gráficas e de linha de comando.

Alguns clientes de grande escala adotaram uma abordagem mais simples, configurando scripts básicos nos hosts para colocar os bancos de dados no modo de backup em um momento específico, em preparação para um snapshot agendado. Por exemplo, programe o comando `alter database begin backup` às 23:58, `alter database end backup` às 00:02 e, em seguida, programe instantâneos diretamente no sistema de armazenamento à meia-noite. O resultado é uma estratégia de backup simples e altamente dimensionável que não requer software ou licenças externos.

Layout de dados

O layout mais simples é isolar datafiles em um ou mais volumes dedicados. Eles devem ser não contaminados por qualquer outro tipo de arquivo. Isso é para garantir que os volumes de arquivo de dados possam ser restaurados rapidamente através de uma operação SnapRestore sem destruir um log refazer importante, controlfile ou log de arquivo.

A SAN tem requisitos semelhantes para isolamento de arquivos de dados dentro de volumes dedicados. Com um sistema operacional como o Microsoft Windows, um único volume pode conter vários LUNs de arquivo de dados, cada um com um sistema de arquivos NTFS. Com outros sistemas operacionais, geralmente há um gerenciador de volumes lógico. Por exemplo, com o Oracle ASM, a opção mais simples seria limitar os LUNs de um grupo de discos ASM a um único volume que pode ser feito backup e restaurado como uma unidade. Se forem necessários volumes adicionais por motivos de gerenciamento de performance ou capacidade, a criação de um grupo de discos adicional no novo volume resultará em um gerenciamento mais simples.

Se essas diretrizes forem seguidas, os snapshots poderão ser agendados diretamente no sistema de storage sem a necessidade de realizar um snapshot de grupo de consistência. A razão é que os backups Oracle não exigem que os datafiles sejam copiados ao mesmo tempo. O procedimento de backup on-line foi projetado para permitir que os arquivos de dados continuem sendo atualizados, pois são transmitidos lentamente para a fita ao longo de horas.

Uma complicação surge em situações como o uso de um grupo de discos ASM que é distribuído entre volumes. Nesses casos, um cg-snapshot deve ser executado para garantir que os metadados ASM sejam consistentes em todos os volumes constituintes.

Atenção: Verifique se o ASM `spfile` e `passwd` os arquivos não estão no grupo de discos que hospeda os arquivos de dados. Isso interfere na capacidade de restaurar seletivamente datafiles e apenas datafiles.

Procedimento de recuperação local – NFS

Este procedimento pode ser conduzido manualmente ou através de uma aplicação como o SnapCenter. O procedimento básico é o seguinte:

1. Encerre o banco de dados.
2. Recupere o(s) volume(s) de arquivo de dados para o instantâneo imediatamente antes do ponto de restauração desejado.
3. Reproduza registos de arquivo até ao ponto pretendido.
4. Repita os logs atuais de refazer se a recuperação completa for desejada.

Este procedimento pressupõe que os registos de arquivo desejados ainda estão presentes no sistema de ficheiros ativo. Se não estiverem, os logs do arquivo devem ser restaurados ou rman/sqlplus podem ser direcionados para os dados no diretório instantâneo.

Além disso, para bancos de dados menores, os arquivos de dados podem ser recuperados por um usuário final diretamente `.snapshot` do diretório sem a ajuda de ferramentas de automação ou administradores de armazenamento para executar um `snaprestore` comando.

Procedimento de recuperação local – SAN

Este procedimento pode ser conduzido manualmente ou através de uma aplicação como o SnapCenter. O procedimento básico é o seguinte:

1. Encerre o banco de dados.
2. Quiesce o(s) grupo(s) de discos que hospedam os arquivos de dados. O procedimento varia consoante o gestor de volume lógico escolhido. Com ASM, o processo requer a desmontagem do grupo de discos. Com o Linux, os sistemas de arquivos devem ser desmontados e os volumes lógicos e grupos de volumes devem ser desativados. O objetivo é parar todas as atualizações no grupo de volume alvo a serem restauradas.
3. Restaure os grupos de discos de arquivo de dados para o instantâneo imediatamente antes do ponto de restauração desejado.
4. Reative os grupos de discos recentemente restaurados.
5. Reproduza registos de arquivo até ao ponto pretendido.
6. Repita todos os logs de refazer se a recuperação completa for desejada.

Este procedimento pressupõe que os registos de arquivo desejados ainda estão presentes no sistema de ficheiros ativo. Se não estiverem, os registos de arquivo devem ser restaurados colocando os LUNs de registo de arquivo offline e executando uma restauração. Este também é um exemplo no qual dividir os logs de arquivo em volumes dedicados é útil. Se os logs de arquivo compartilharem um grupo de volumes com os logs de refazer, os logs de refazer devem ser copiados em outro lugar antes da restauração do conjunto geral de LUNs. Esta etapa impede a perda dessas transações finais registradas.

Backups otimizados para Storage Snapshot

Backup e recuperação baseados em snapshot se tornaram ainda mais simples quando o Oracle 12c foi lançado porque não há necessidade de colocar um banco de dados no modo hot backup. O resultado é a capacidade de agendar backups baseados em snapshot diretamente em um sistema de storage e ainda preservar a capacidade de executar recuperação completa ou pontual.

Embora o procedimento de recuperação de hot backup seja mais familiar aos DBAs, há muito tempo foi possível usar snapshots que não foram criados enquanto o banco de dados estava no modo hot backup. Etapas manuais extras foram necessárias com o Oracle 10gi e 11gi durante a recuperação para tornar o banco de dados consistente. Com o Oracle 12ci, `sqlplus` e `rman` conter a lógica extra para reproduzir logs de arquivo em backups de arquivos de dados que não estavam no modo de backup ativo.

Como discutido anteriormente, a recuperação de um hot backup baseado em snapshot requer dois conjuntos de dados:

- Um instantâneo dos arquivos de dados criados no modo de backup
- Os logs de arquivo gerados enquanto os datafiles estavam no modo hot backup

Durante a recuperação, o banco de dados lê metadados dos arquivos de dados para selecionar os logs de arquivo necessários para recuperação.

A recuperação otimizada para snapshot de storage requer conjuntos de dados ligeiramente diferentes para alcançar os mesmos resultados:

- Um instantâneo dos arquivos de dados, além de um método para identificar a hora em que o snapshot foi criado
- Arquive logs do tempo do ponto de verificação mais recente do arquivo de dados até a hora exata do instantâneo

Durante a recuperação, o banco de dados lê metadados dos arquivos de dados para identificar o Registro de arquivo mais antigo necessário. A recuperação completa ou pontual pode ser realizada. Ao executar uma recuperação pontual, é fundamental saber o tempo do snapshot dos arquivos de dados. O ponto de recuperação especificado deve ser após o tempo de criação dos instantâneos. A NetApp recomenda adicionar pelo menos alguns minutos à hora do instantâneo para contabilizar a variação do relógio.

Para obter detalhes completos, consulte a documentação da Oracle sobre o tópico "recuperação usando Otimização de Snapshot de armazenamento" disponível em várias versões da documentação do Oracle 12c. Além disso, consulte Oracle Document ID Doc ID 604683,1 sobre o suporte a snapshots de terceiros da Oracle.

Layout de dados

O layout mais simples é isolar os arquivos de dados em um ou mais volumes dedicados. Eles devem ser não contaminados por qualquer outro tipo de arquivo. Isso é para garantir que os volumes de arquivo de dados possam ser restaurados rapidamente com uma operação SnapRestore sem destruir um log de refazer importante, controlfile ou log de arquivo.

A SAN tem requisitos semelhantes para isolamento de arquivos de dados dentro de volumes dedicados. Com um sistema operacional como o Microsoft Windows, um único volume pode conter vários LUNs de arquivo de dados, cada um com um sistema de arquivos NTFS. Com outros sistemas operacionais, geralmente há um gerenciador de volume lógico também. Por exemplo, com o Oracle ASM, a opção mais simples seria limitar grupos de discos a um único volume que pode ser feito backup e restaurado como uma unidade. Se forem necessários volumes adicionais por motivos de gerenciamento de performance ou capacidade, criar um grupo de discos adicional no novo volume resulta em gerenciamento mais fácil.

Se essas diretrizes forem seguidas, os snapshots poderão ser agendados diretamente no ONTAP sem a necessidade de realizar um snapshot de grupo de consistência. O motivo é que backups otimizados para snapshot não exigem que sejam feitos backup de dados ao mesmo tempo.

Uma complicação surge em situações como um grupo de discos ASM que é distribuído entre volumes. Nesses casos, um cg-snapshot deve ser executado para garantir que os metadados ASM sejam consistentes em todos os volumes constituintes.

[Nota]Verifique se os arquivos ASM spfile e passwd não estão no grupo de discos que hospeda os arquivos de dados. Isso interfere na capacidade de restaurar seletivamente datafiles e apenas datafiles.

Procedimento de recuperação local – NFS

Este procedimento pode ser conduzido manualmente ou através de uma aplicação como o SnapCenter. O procedimento básico é o seguinte:

1. Encerre o banco de dados.
2. Recupere o(s) volume(s) de arquivo de dados para o instantâneo imediatamente antes do ponto de restauração desejado.
3. Reproduza registos de arquivo até ao ponto pretendido.

Este procedimento pressupõe que os registos de arquivo desejados ainda estão presentes no sistema de ficheiros ativo. Se não estiverem, os registos de arquivo têm de ser restaurados ou `rman` `sqlplus` podem ser direcionados para os dados no `.snapshot` diretório.

Além disso, para bancos de dados menores, os arquivos de dados podem ser recuperados por um usuário final diretamente `.snapshot` do diretório sem a ajuda de ferramentas de automação ou um administrador de armazenamento para executar um comando SnapRestore.

Procedimento de recuperação local – SAN

Este procedimento pode ser conduzido manualmente ou através de uma aplicação como o SnapCenter. O procedimento básico é o seguinte:

1. Encerre o banco de dados.
2. Quiesce o(s) grupo(s) de discos que hospedam os arquivos de dados. O procedimento varia consoante o gestor de volume lógico escolhido. Com ASM, o processo requer a desmontagem do grupo de discos. Com o Linux, os sistemas de arquivos devem ser desmontados e os volumes lógicos e grupos de volumes são desativados. O objetivo é parar todas as atualizações no grupo de volume alvo a serem restauradas.
3. Restaure os grupos de discos de arquivo de dados para o instantâneo imediatamente antes do ponto de restauração desejado.
4. Reative os grupos de discos recentemente restaurados.
5. Reproduza registos de arquivo até ao ponto pretendido.

Este procedimento pressupõe que os registos de arquivo desejados ainda estão presentes no sistema de ficheiros ativo. Se não estiverem, os registos de arquivo devem ser restaurados colocando os LUNs de registo de arquivo offline e executando uma restauração. Este também é um exemplo no qual dividir os logs de arquivo em volumes dedicados é útil. Se os logs do arquivo compartilharem um grupo de volumes com os logs de refazer, os logs de refazer devem ser copiados em outro lugar antes da restauração do conjunto geral de LUNs para evitar perder as transações registradas finais.

Exemplo de recuperação completa

Suponha que os arquivos de dados foram corrompidos ou destruídos e a recuperação completa é necessária. O procedimento para o fazer é o seguinte:

```
[oracle@host1 ~]$ sqlplus / as sysdba
Connected to an idle instance.
SQL> startup mount;
ORACLE instance started.
Total System Global Area 1610612736 bytes
Fixed Size                  2924928 bytes
Variable Size            1040191104 bytes
Database Buffers          553648128 bytes
Redo Buffers               13848576 bytes
Database mounted.
SQL> recover automatic;
Media recovery complete.
SQL> alter database open;
Database altered.
SQL>
```

Exemplo de recuperação pontual

Todo o procedimento de recuperação é um único comando: `recover automatic`.

Se a recuperação pontual for necessária, o carimbo de data/hora do(s) instantâneo(s) deve(m) ser conhecido(s) e pode(m) ser identificado(s) da seguinte forma:

```
Cluster01::> snapshot show -vserver vserver1 -volume NTAP_oradata -fields
create-time
vserver   volume        snapshot   create-time
--------  ------------  ---------  ------------------------
vserver1  NTAP_oradata  my-backup  Thu Mar 09 10:10:06 2017
```

A hora de criação do instantâneo é listada como 9th de Março e 10:10:06. Para estar seguro, um minuto é adicionado à hora do instantâneo:

```
[oracle@host1 ~]$ sqlplus / as sysdba
Connected to an idle instance.
SQL> startup mount;
ORACLE instance started.
Total System Global Area 1610612736 bytes
Fixed Size                  2924928 bytes
Variable Size            1040191104 bytes
Database Buffers          553648128 bytes
Redo Buffers               13848576 bytes
Database mounted.
SQL> recover database until time '09-MAR-2017 10:44:15' snapshot time '09-
MAR-2017 10:11:00';
```

A recuperação agora é iniciada. Ele especificou um tempo instantâneo de 10:11:00, um minuto após o tempo gravado para contabilizar a possível variação do relógio e um tempo de recuperação alvo de 10:44. Em seguida, sqlplus solicita os logs de arquivo necessários para alcançar o tempo de recuperação desejado de 10:44.

```
ORA-00279: change 551760 generated at 03/09/2017 05:06:07 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_31_930813377.dbf
ORA-00280: change 551760 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 552566 generated at 03/09/2017 05:08:09 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_32_930813377.dbf
ORA-00280: change 552566 for thread 1 is in sequence #32
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 553045 generated at 03/09/2017 05:10:12 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_33_930813377.dbf
ORA-00280: change 553045 for thread 1 is in sequence #33
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 753229 generated at 03/09/2017 05:15:58 needed for
thread 1
ORA-00289: suggestion : /oralogs_nfs/arch/1_34_930813377.dbf
ORA-00280: change 753229 for thread 1 is in sequence #34
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
Log applied.
Media recovery complete.
SQL> alter database open resetlogs;
Database altered.
SQL>
```

A recuperação completa de um banco de dados usando snapshots usando o `recover automatic` comando não requer licenciamento específico, mas a recuperação pontual usando `snapshot time` requer a licença Oracle Advanced Compression.

Ferramentas de gerenciamento e automação de banco de dados

O principal valor do ONTAP em um ambiente de banco de dados Oracle vem das principais tecnologias da ONTAP, como cópias Snapshot instantâneas, replicação simples do SnapMirror e criação eficiente de volumes do FlexClone.

Em alguns casos, a configuração simples desses principais recursos diretamente no ONTAP atende aos requisitos, mas as necessidades mais complicadas exigem uma camada de orquestração.

SnapCenter

O SnapCenter é o principal produto de proteção de dados da NetApp. Em um nível muito baixo, ele é

semelhante aos produtos SnapManager em termos de como executa backups de bancos de dados. No entanto, ele foi criado do zero para fornecer um painel único para gerenciamento de proteção de dados em sistemas de storage da NetApp.

O SnapCenter inclui funções básicas, como backups e restaurações baseados em snapshot, replicação SnapMirror e SnapVault e outros recursos necessários para operar em escala para empresas de grande porte. Esses recursos avançados incluem funcionalidade de controle de acesso baseado em funções (RBAC) expandida, APIs RESTful para integração com produtos de orquestração de terceiros, gerenciamento central sem interrupções de plug-ins do SnapCenter em hosts de banco de dados e uma interface de usuário projetada para ambientes de escala de nuvem.

DESCANSO

O ONTAP também contém um conjunto de APIs RESTful. Isso permite que fornecedores terceirizados de 3rd criem proteção de dados e outros aplicativos de gerenciamento com profunda integração com o ONTAP. Além disso, a API RESTful é fácil de consumir por clientes que desejam criar seus próprios fluxos de trabalho e utilitários de automação.

Recuperação de desastres Oracle

Visão geral

Recuperação de desastres refere-se à restauração de serviços de dados após um evento catastrófico, como um incêndio que destrói um sistema de storage ou até mesmo um local inteiro.

Esta documentação substitui os relatórios técnicos publicados anteriormente *TR-4591: Oracle Data Protection* e *TR-4592: Oracle no MetroCluster.*

A recuperação de desastres pode ser realizada por replicação simples de dados usando o SnapMirror, é claro, com muitos clientes atualizando réplicas espelhadas quantas vezes por hora.

Para a maioria dos clientes, a recuperação de desastres requer mais do que apenas a posse de uma cópia remota de dados. Isso exige a capacidade de usar esses dados rapidamente. A NetApp oferece duas tecnologias que atendem a essa necessidade: MetroCluster e SnapMirror ative Sync

O MetroCluster se refere ao ONTAP em uma configuração de hardware que inclui armazenamento de baixo nível espelhado e vários recursos adicionais. Soluções integradas, como o MetroCluster, simplificam as infraestruturas de virtualização, aplicações e banco de dados complicadas e com escalabilidade horizontal atuais. Ele substitui vários produtos e estratégias de proteção de dados externos por um storage array simples e central. Ele também fornece backup integrado, recuperação, recuperação de desastres e alta disponibilidade (HA) em um único sistema de storage em cluster.

A sincronização ativa do SnapMirror (SM-as) é baseada no SnapMirror síncrono. Com o MetroCluster, cada controlador ONTAP é responsável por replicar os dados da unidade para um local remoto. Com o SnapMirror active Sync, você tem essencialmente dois sistemas ONTAP diferentes que mantêm cópias independentes dos seus dados LUN, mas cooperam para apresentar uma única instância desse LUN. Do ponto de vista do host, é uma única entidade LUN.

Comparação de SM-as e MCC

O SM-as e o MetroCluster são semelhantes na funcionalidade geral, mas há diferenças importantes na maneira como a replicação RPO-0 foi implementada e como ela é gerenciada. O SnapMirror assíncrono e

síncrono também pode ser usado como parte de um plano de recuperação de desastres, mas não foram desenvolvidos como tecnologias de repliação de HA.

- Uma configuração do MetroCluster é mais como um cluster integrado com nós distribuídos entre locais. O SM-as se comporta como dois clusters independentes que estão cooperando no fornecimento de LUNs replicados de forma síncrona RPO igual a 0.
- Os dados em uma configuração do MetroCluster só podem ser acessados de um determinado site a qualquer momento. Uma segunda cópia dos dados está presente no site oposto, mas os dados são passivos. Ele não pode ser acessado sem um failover do sistema de storage.
- O espelhamento de desempenho MetroCluster e SM-as ocorre em níveis diferentes. O espelhamento MetroCluster é executado na camada RAID. Os dados de baixo nível são armazenados em um formato espelhado usando SyncMirror. O uso do espelhamento é praticamente invisível nas camadas de LUN, volume e protocolo.
- Em contraste, o espelhamento SM-as ocorre na camada de protocolo. Os dois clusters são, no geral, clusters independentes. Depois que as duas cópias dos dados estiverem sincronizadas, os dois clusters só precisam espelhar gravações. Quando ocorre uma gravação em um cluster, ela é replicada para o outro cluster. A gravação só é reconhecida para o host quando a gravação for concluída em ambos os sites. Além desse comportamento de divisão de protocolo, os dois clusters são, de outra forma, clusters ONTAP normais.
- A função principal do MetroCluster é a replicação em grande escala. É possível replicar um array inteiro com RPO igual a 0 e rto quase zero. Isso simplifica o processo de failover porque há apenas uma "coisa" a fazer failover e é dimensionado extremamente bem em termos de capacidade e IOPS.
- Um dos principais casos de uso para SM-as é a replicação granular. Às vezes, você não quer replicar todos os dados como uma única unidade ou precisa ser capaz de falhar seletivamente em determinados workloads.
- Outro importante caso de uso para SM-as é para operações ativas-ativas, em que você deseja que cópias totalmente utilizáveis de dados estejam disponíveis em dois clusters diferentes localizados em dois locais diferentes com características de desempenho idênticas e, se desejado, não é necessário estender a SAN entre locais. Você pode ter suas aplicações já em execução em ambos os locais, o que reduz o rto geral durante operações de failover.

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.

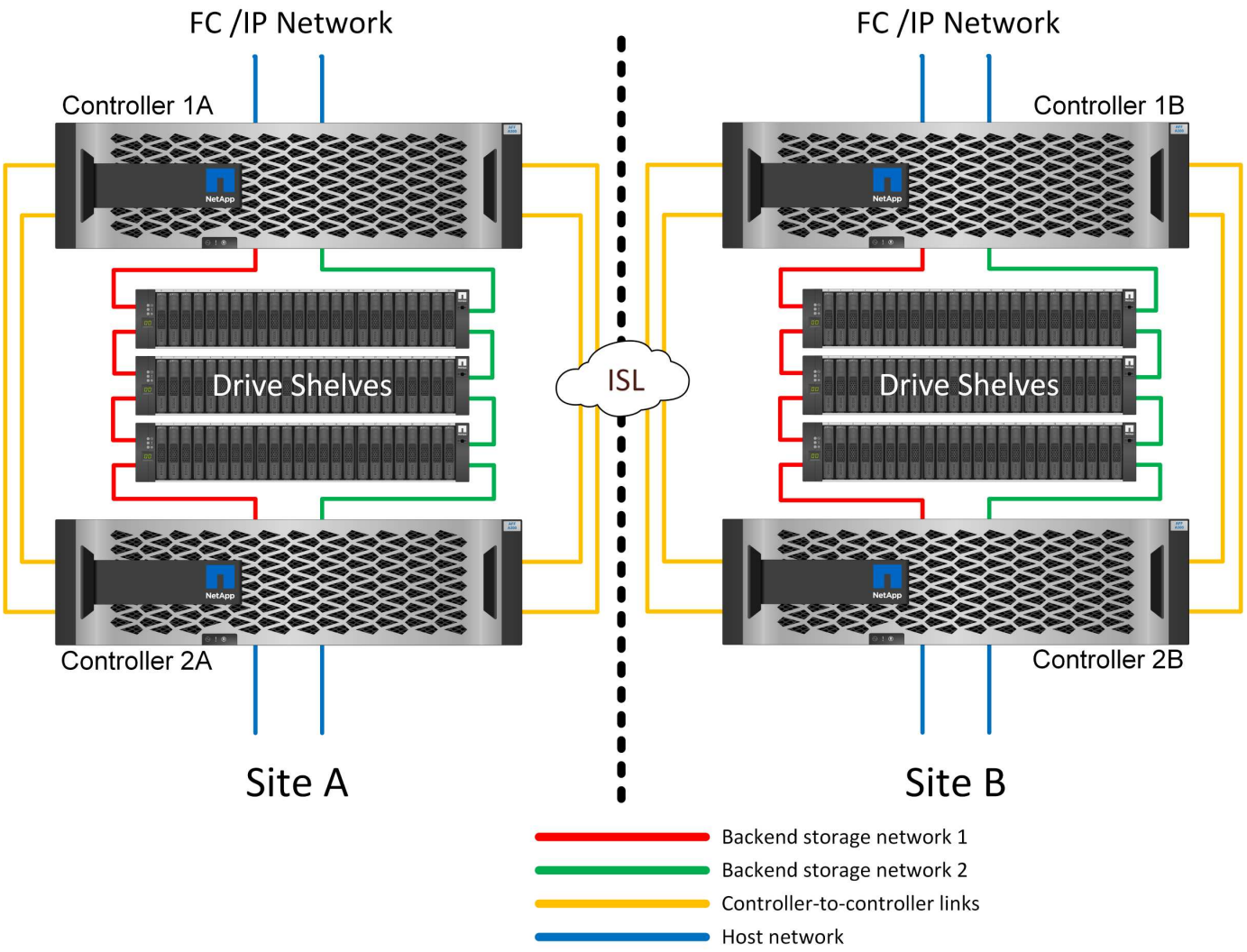
IP MetroCluster

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.

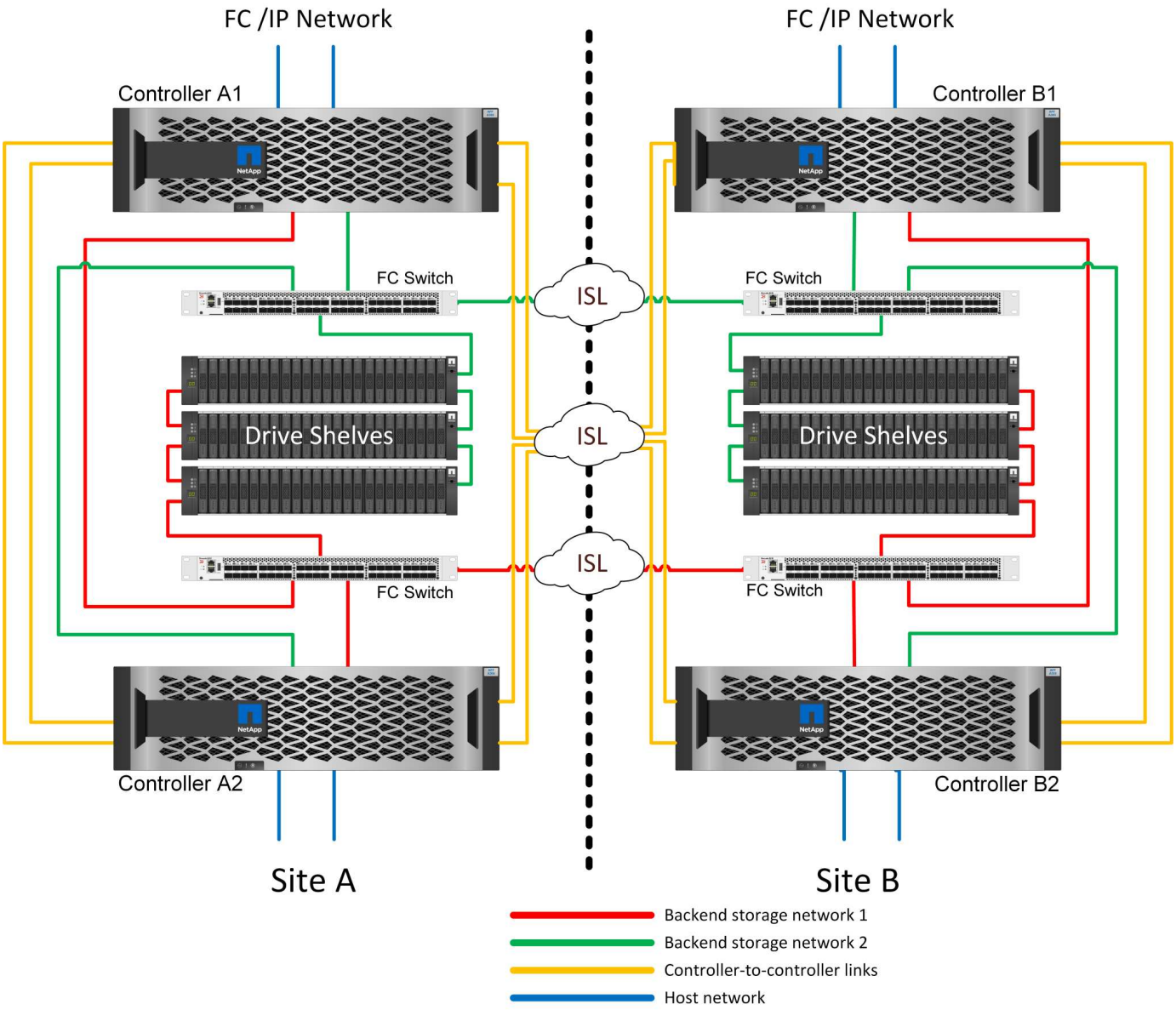
A arquitetura também é mais simples porque as únicas conexões entre locais são para os controladores. No Metroclusters FC SAN conetados, um controlador grava diretamente nas unidades no local oposto e, portanto, requer conexões, switches e bridges SAN adicionais. Em contraste, um controlador em uma configuração 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.

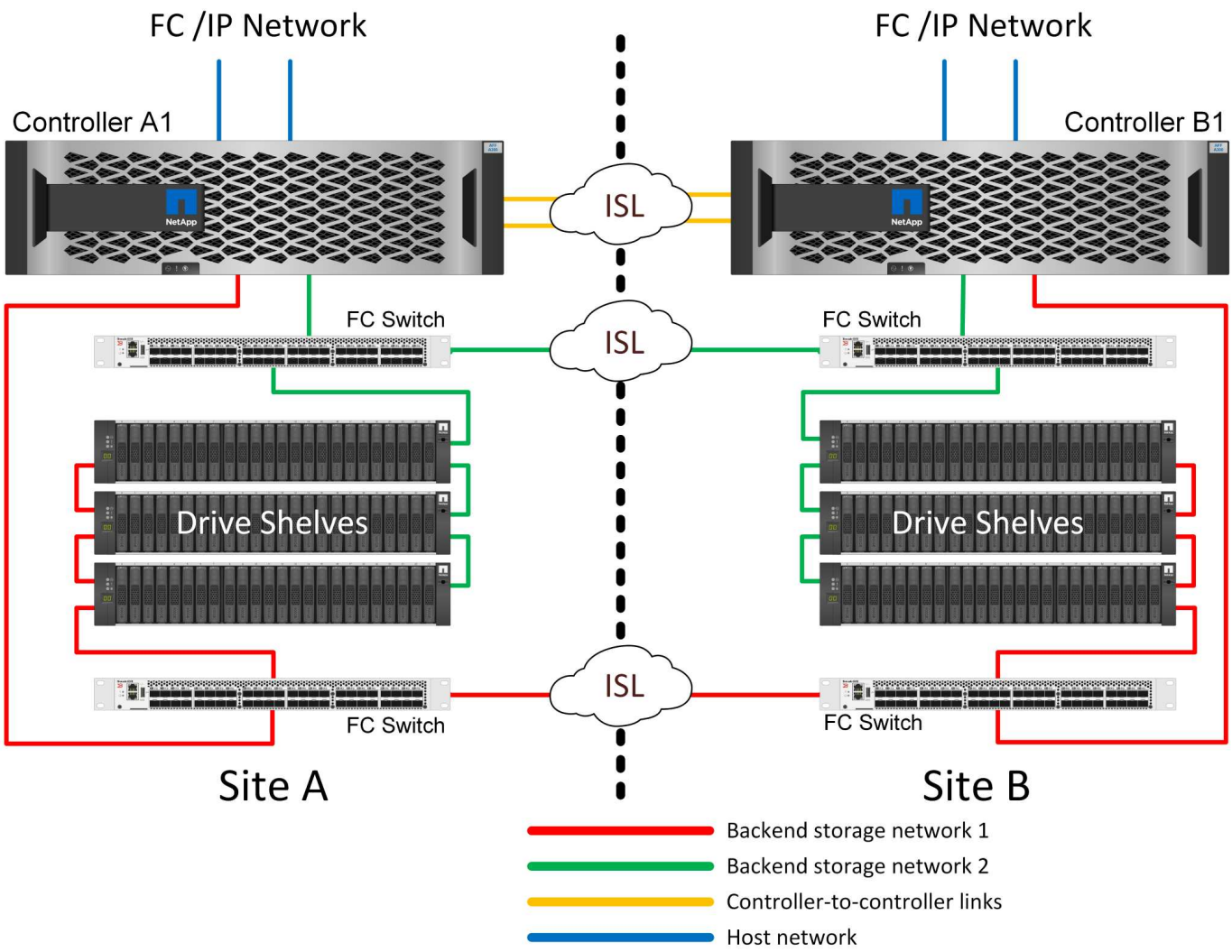


Algumas infraestruturas multisite não foram desenvolvidas para operações ativas-ativas, mas são usadas mais como local principal e local de recuperação de desastres. Nesta situação, uma opção MetroCluster de par de HA é geralmente preferível pelas seguintes razões:

- 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

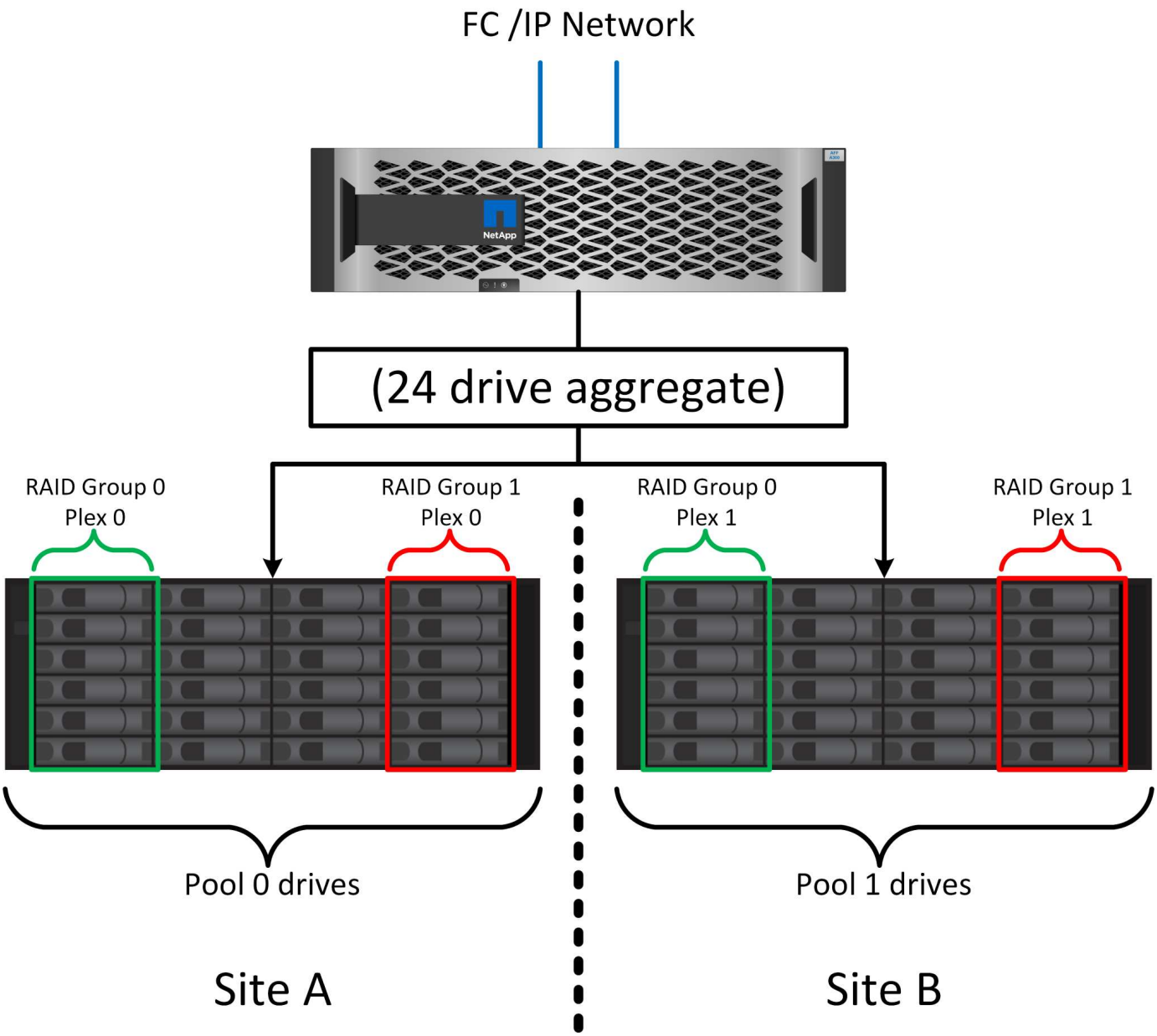
A MetroCluster estende a proteção de dados da NVRAM das seguintes maneiras:

- 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.

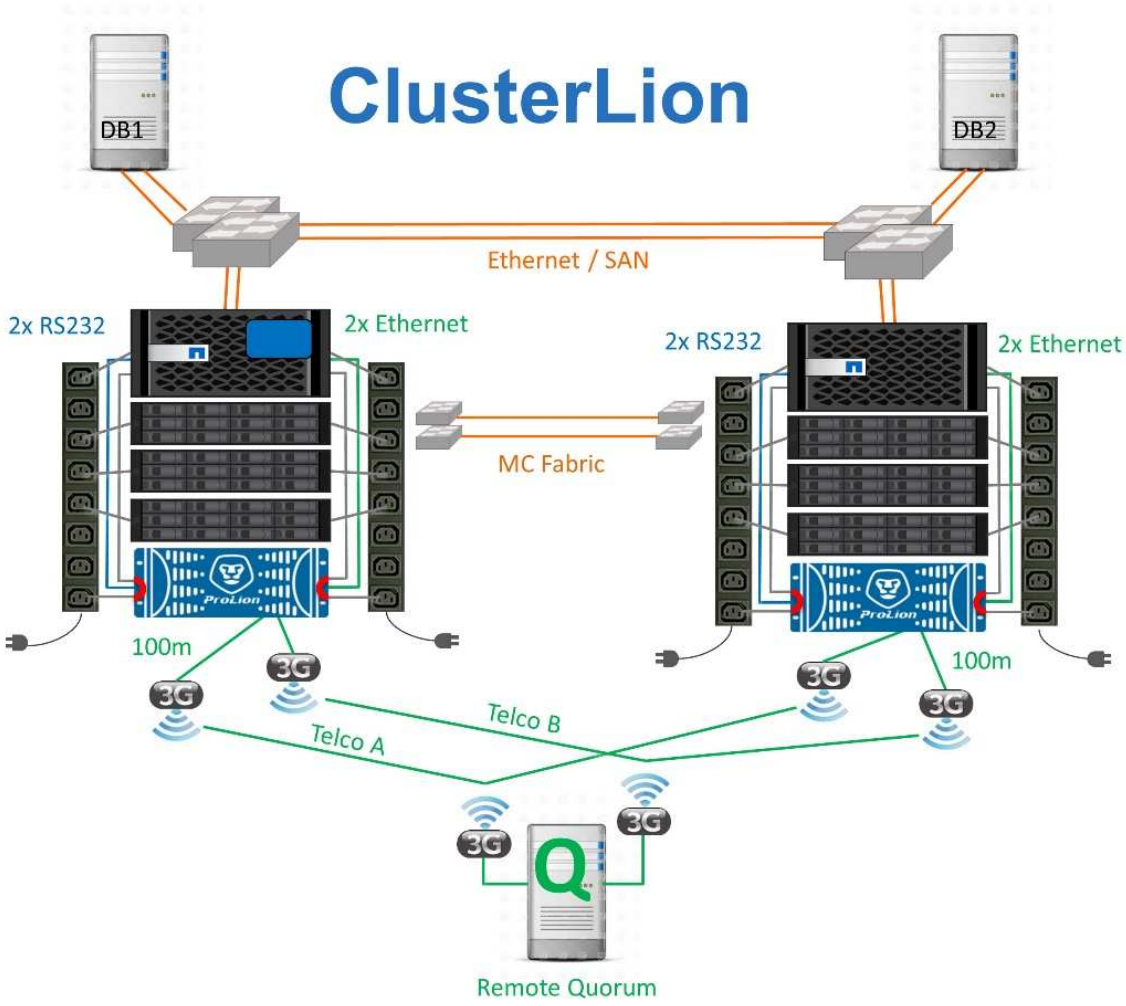
Um MetroCluster conectado à malha tem acesso direto aos dispositivos de storage no local oposto. Isso permite que um controlador MetroCluster monitore a integridade dos outros controladores lendo dados de batimentos cardíacos das unidades. Isso permite que um controlador reconheça a falha de outro controlador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

Sincronização ativa do SnapMirror

Visão geral

O SnapMirror ative Sync permite que você crie ambientes de banco de dados Oracle de alta disponibilidade, onde LUNs estão disponíveis em dois clusters de storage diferentes.

Com a sincronização ativa do SnapMirror, não há cópia "primária" e "secundária" dos dados. Cada cluster pode fornecer leitura de e/S de sua cópia local dos dados, e cada cluster replicará uma gravação para seu parceiro. O resultado é um comportamento de IO simétrico.

Entre outras opções, isso permite que você execute o Oracle RAC como um cluster estendido com instâncias operacionais em ambos os sites. Como alternativa, você pode criar clusters de banco de dados ativo-passivo RPO igual a 0 em que bancos de dados de instância única podem ser movidos por locais durante uma falha no local. Esse processo pode ser automatizado por meio de produtos como pacemaker ou VMware HA. A base para todas essas opções é a replicação síncrona gerenciada pela sincronização ativa do SnapMirror.

Replicação síncrona

Em operação normal, o SnapMirror active Sync fornece réplica síncrona RPO igual a 0 em todos os momentos, com uma exceção. Se os dados não puderem ser replicados, o ONTAP cumprirá o requisito de replicar dados e retomar a distribuição de I/o em um local, enquanto os LUNs no outro local ficam offline.

Hardware de storage

Ao contrário de outras soluções de recuperação de desastres de storage, o SnapMirror active Sync oferece flexibilidade assimétrica de plataforma. O hardware em cada local não precisa ser idêntico. Esse recurso permite dimensionar corretamente o hardware usado para suportar a sincronização ativa do SnapMirror. O sistema de storage remoto pode ser idêntico ao local principal se precisar dar suporte a uma carga de trabalho de produção completa, mas se um desastre resultar em e/S reduzida, do que um sistema menor no local remoto pode ser mais econômico.

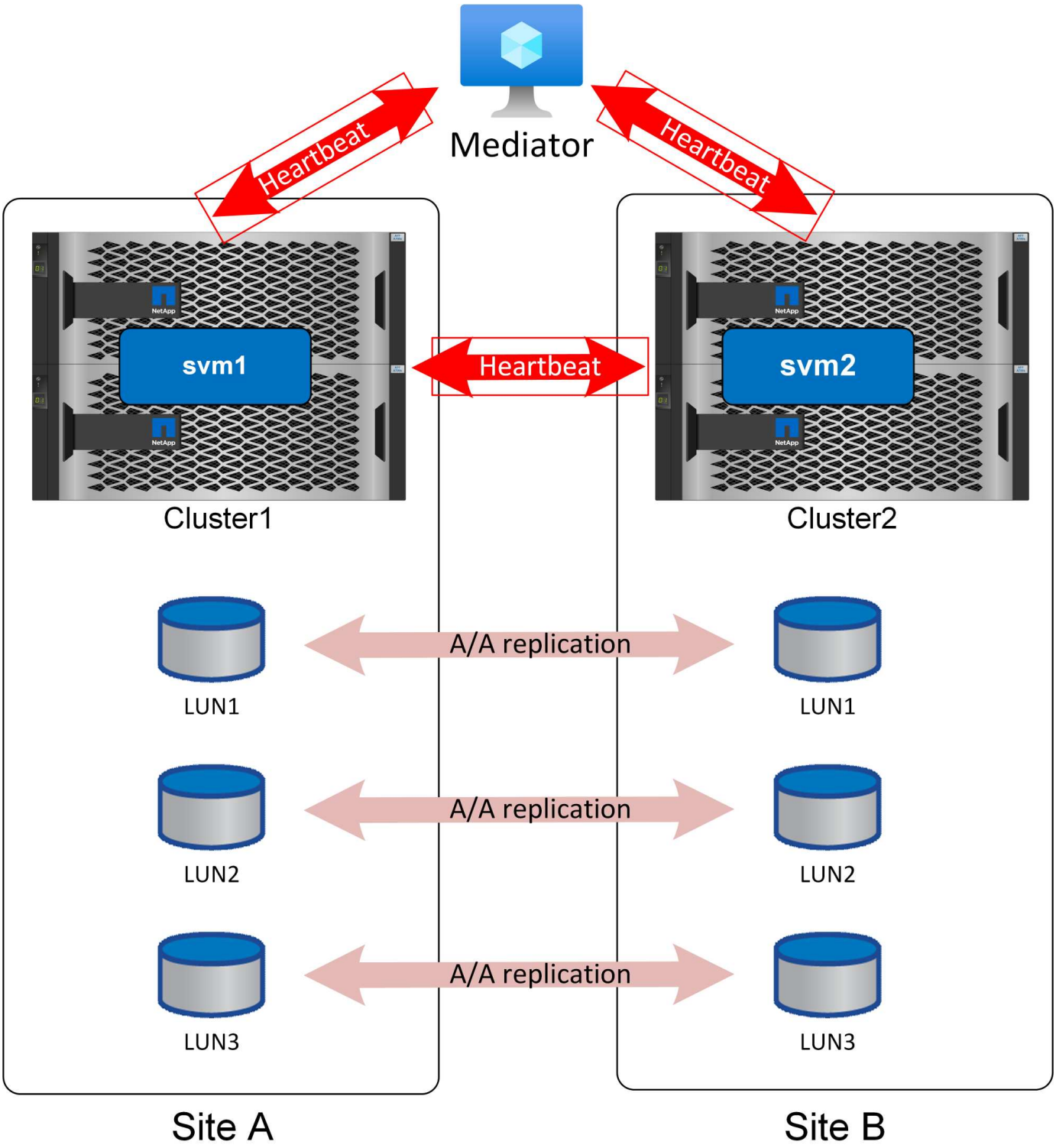
Mediador do ONTAP

O Mediador ONTAP é um aplicativo de software que é baixado do suporte do NetApp e normalmente é implantado em uma pequena máquina virtual. O Mediador ONTAP não é um tiebreaker quando usado com a sincronização ativa do SnapMirror. É um canal de comunicação alternativo para os dois clusters que participam da replicação de sincronização ativa do SnapMirror. As operações automatizadas são orientadas pelo ONTAP com base nas respostas recebidas do parceiro por meio de conexões diretas e por meio do mediador.

ONTAP Mediador

O mediador é necessário para automatizar o failover com segurança. Idealmente, ele seria colocado em um local 3rd independente, mas ainda pode funcionar para a maioria das necessidades se colocasse em um dos clusters que participam da replicação.

O mediador não é propriamente um desempate, embora essa seja efetivamente a função que desempenha. O mediador ajuda a determinar o estado dos nós do cluster e auxilia no processo de comutação automática em caso de falha de um dos sites. O Mediator não transfere dados sob nenhuma circunstância.



O desafio nº 1 com failover automatizado é o problema de split-brain, e esse problema surge se os dois locais perderem a conetividade entre si. O que deve acontecer? Você não quer que dois sites diferentes se designem como as cópias sobreviventes dos dados, mas como um único site pode dizer a diferença entre a perda real do site oposto e a incapacidade de se comunicar com o site oposto?

É aqui que o mediador entra na imagem. Se for colocado em um site 3rd e cada site tiver uma conexão de rede separada com esse site, então você terá um caminho adicional para cada site validar a integridade do outro. Olhe para a imagem acima novamente e considere os seguintes cenários.

- O que acontece se o mediador falhar ou não estiver acessível a partir de um ou de ambos os sites?
  - Os dois clusters ainda podem se comunicar entre si pelo mesmo link usado para serviços de replicação.
  - Os dados ainda são servidos com proteção RPO igual a 0
- O que acontece se o Site A falhar?
  - O local B verá ambos os canais de comunicação diminuírem.
  - O local B assumirá os serviços de dados, mas sem o espelhamento RPO igual a 0
- O que acontece se o local B falhar?
  - O local A verá ambos os canais de comunicação diminuírem.
  - O local A assumirá os serviços de dados, mas sem o espelhamento do RPO igual a 0

Há um outro cenário a considerar: Perda do link de replicação de dados. Se o link de replicação entre locais for perdido, o espelhamento RPO 0 obviamente será impossível. O que deve acontecer então?

Isso é controlado pelo status do site preferido. Em uma relação SM-as, um dos locais é secundário ao outro. Isso não tem efeito nas operações normais, e todo o acesso aos dados é simétrico, mas se a replicação for interrompida, o empate terá que ser quebrado para retomar as operações. O resultado é que o local preferido continuará as operações sem espelhamento, e o local secundário interromperá o processamento de e/S até que a comunicação de replicação seja restaurada.

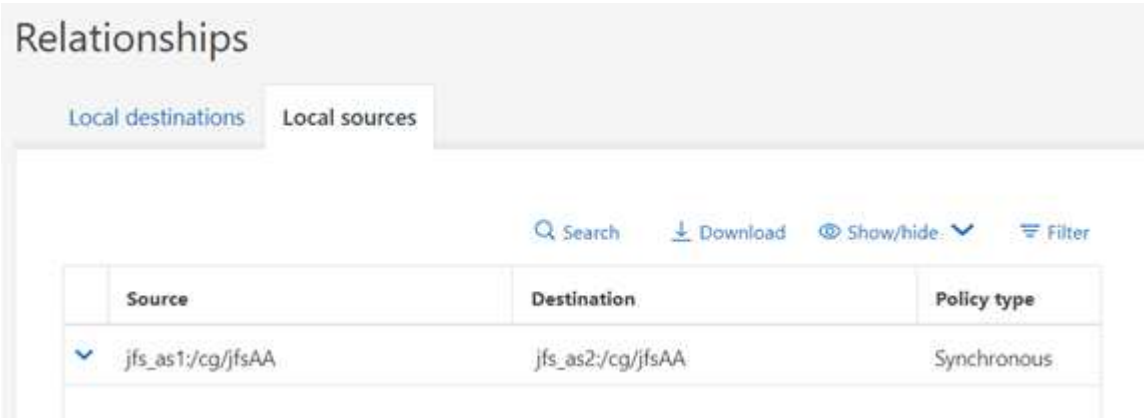
Site preferido para sincronização ativa do SnapMirror

O comportamento de sincronização ativa do SnapMirror é simétrico, com uma exceção importante - configuração de site preferida.

A sincronização ativa do SnapMirror considerará um site a "fonte" e o outro o "destino". Isso implica uma relação de replicação unidirecional, mas isso não se aplica ao comportamento de IO. A replicação é bidirecional e simétrica, e os tempos de resposta de e/S são os mesmos em ambos os lados do espelho.

A `source` designação é controla o local preferido. Se o link de replicação for perdido, os caminhos de LUN na cópia de origem continuarão a servir dados enquanto os caminhos de LUN na cópia de destino ficarão indisponíveis até que a replicação seja restabelecida e o SnapMirror reinsira um estado síncrono. Os caminhos irão então retomar a veiculação de dados.

A configuração de origem/destino pode ser visualizada através do SystemManager:



Ou na CLI:

```
Cluster2::> snapmirror show -destination-path jfs_as2:/cg/jfsAA

                           Source Path: jfs_as1:/cg/jfsAA
                     Destination Path: jfs_as2:/cg/jfsAA
                    Relationship Type: XDP
                Relationship Group Type: consistencygroup
                    SnapMirror Schedule: -
                 SnapMirror Policy Type: automated-failover-duplex
                      SnapMirror Policy: AutomatedFailOverDuplex
                            Tries Limit: -
                      Throttle (KB/sec): -
                           Mirror State: Snapmirrored
                    Relationship Status: InSync
```

O segredo é que a fonte é o SVM no cluster1. Como mencionado acima, os termos "fonte" e "destino" não descrevem o fluxo de dados replicados. Ambos os sites podem processar uma gravação e replicá-la para o site oposto. Com efeito, ambos os clusters são fontes e destinos. O efeito de designar um cluster como uma fonte simplesmente controla qual cluster sobrevive como um sistema de armazenamento de leitura e gravação se o link de replicação for perdido.

Topologia de rede

Acesso uniforme

Rede de acesso uniforme significa que os hosts são capazes de acessar caminhos em ambos os sites (ou domínios de falha dentro do mesmo site).

Um recurso importante do SM-as é a capacidade de configurar os sistemas de storage para saber onde os hosts estão localizados. Quando você mapeia os LUNs para um determinado host, você pode indicar se eles são ou não proximais a um determinado sistema de armazenamento.

Definições de proximidade

Proximidade refere-se a uma configuração por cluster que indica que um determinado host WWN ou ID de iniciador iSCSI pertence a um host local. É uma segunda etapa opcional para configurar o acesso LUN.

O primeiro passo é a configuração usual do igroup. Cada LUN deve ser mapeado para um grupo que contenha as IDs WWN/iSCSI dos hosts que precisam de acesso a esse LUN. Isso controla qual host tem *access* para um LUN.

A segunda etapa opcional é configurar a proximidade do host. Isso não controla o acesso, ele controla *Priority*.

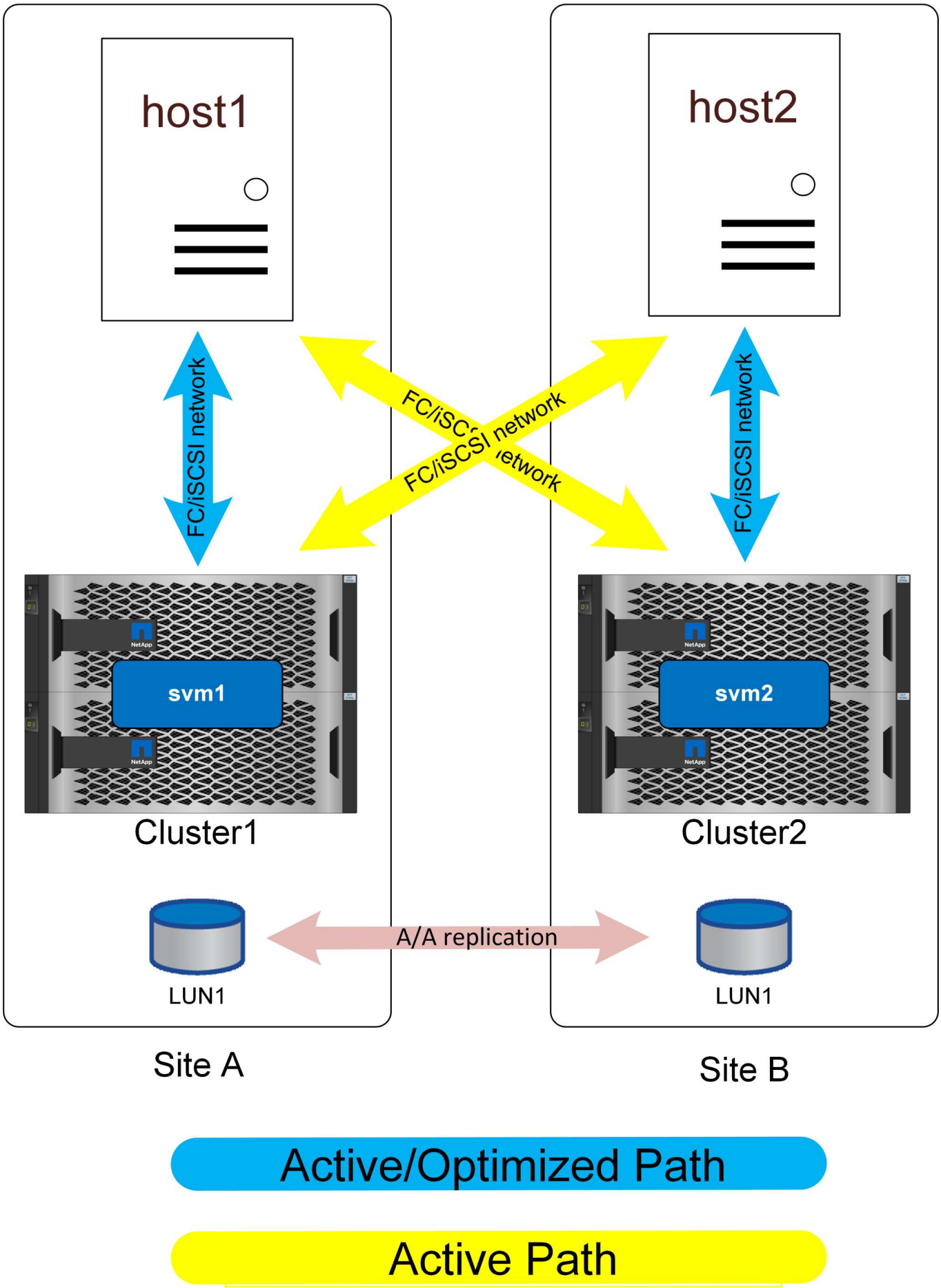
Por exemplo, um host no local A pode ser configurado para acessar um LUN que é protegido pela sincronização ativa do SnapMirror e, como a SAN é estendida entre sites, há caminhos disponíveis para esse LUN usando armazenamento no local A ou armazenamento no local B.

Sem configurações de proximidade, esse host usará ambos os sistemas de storage igualmente porque ambos os sistemas de storage anunciarão caminhos ativos/otimizados. Se a latência da SAN e/ou a largura de banda entre locais for limitada, isso pode não ser desejado e você pode querer garantir que, durante a operação normal, cada host utilize preferencialmente caminhos para o sistema de armazenamento local. Isso é

configurado adicionando o ID WWN/iSCSI do host ao cluster local como um host proximal. Isso pode ser feito na CLI ou no SystemManager.

AFF

Com um sistema AFF, os caminhos aparecerão como mostrado abaixo quando a proximidade do host for configurada.


host1
host2
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
svm1
svm2
Cluster1
Cluster2
A/A replication
LUN1
LUN1
Site A
Site B
Active/Optimized Path
Active Path

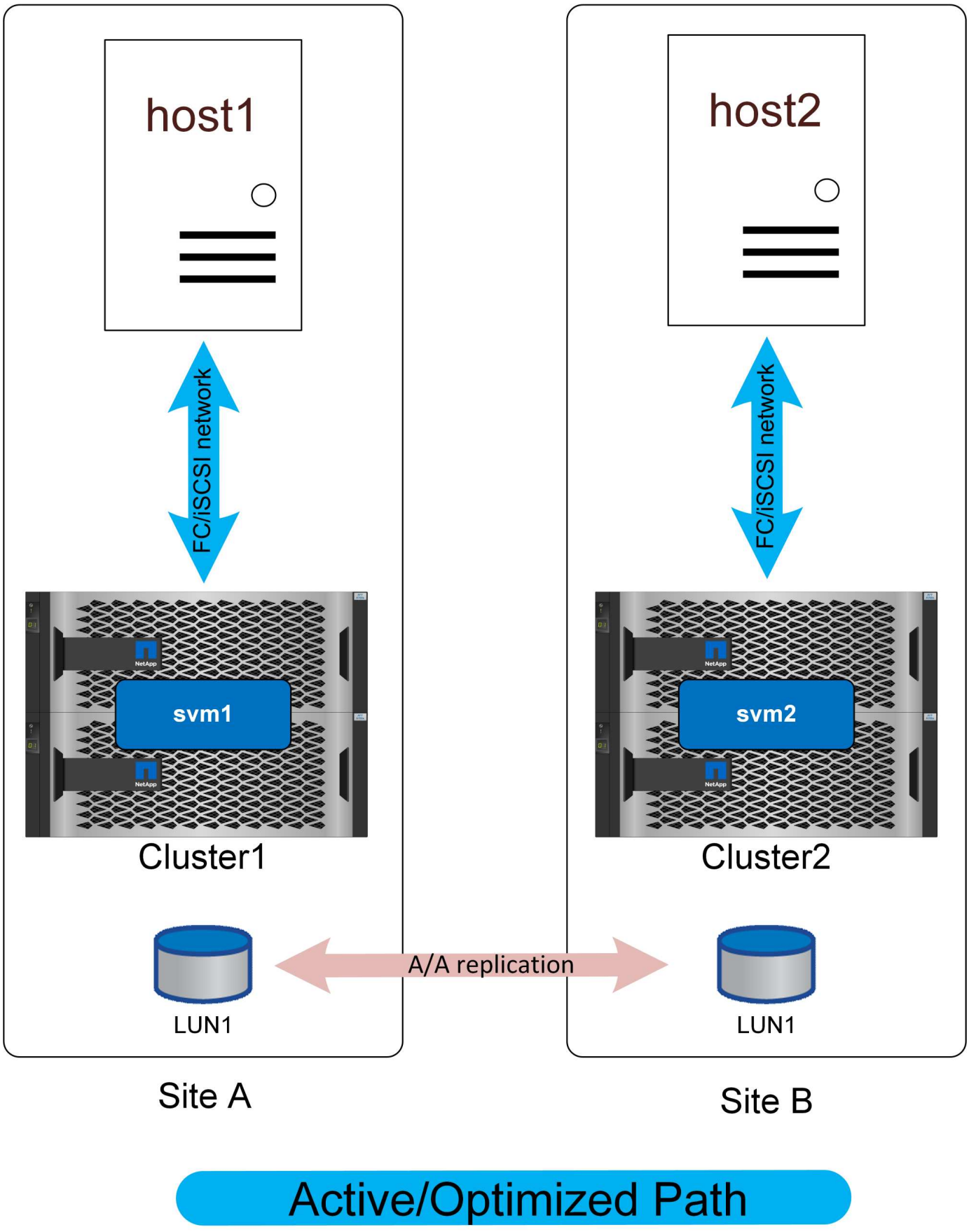
Em operação normal, todo o IO é IO local. Leituras e gravações são atendidas a partir do storage array local. É claro que o write IO também precisará ser replicado pelo controlador local para o sistema remoto antes de ser reconhecido, mas todas as IO de leitura serão atendidas localmente e não incorrerão latência extra ao atravessar o link SAN entre locais.

A única vez que os caminhos não otimizados serão usados é quando todos os caminhos ativos/otimizados forem perdidos. Por exemplo, se todo o array no local perder energia, os hosts no local A ainda poderão acessar caminhos para o array no local B e, portanto, permanecer operacionais, embora estejam com maior latência.

Há caminhos redundantes pelo cluster local que não são mostrados nesses diagramas por uma questão de simplicidade. Os sistemas de storage da ONTAP estão HA, portanto, uma falha da controladora não deve resultar em falha do local. Deve apenas resultar em uma mudança na qual os caminhos locais são usados no site afetado.

ASA

Os sistemas NetApp ASA oferecem multipathing ativo-ativo em todos os caminhos em um cluster. Isso também se aplica às configurações SM-as.



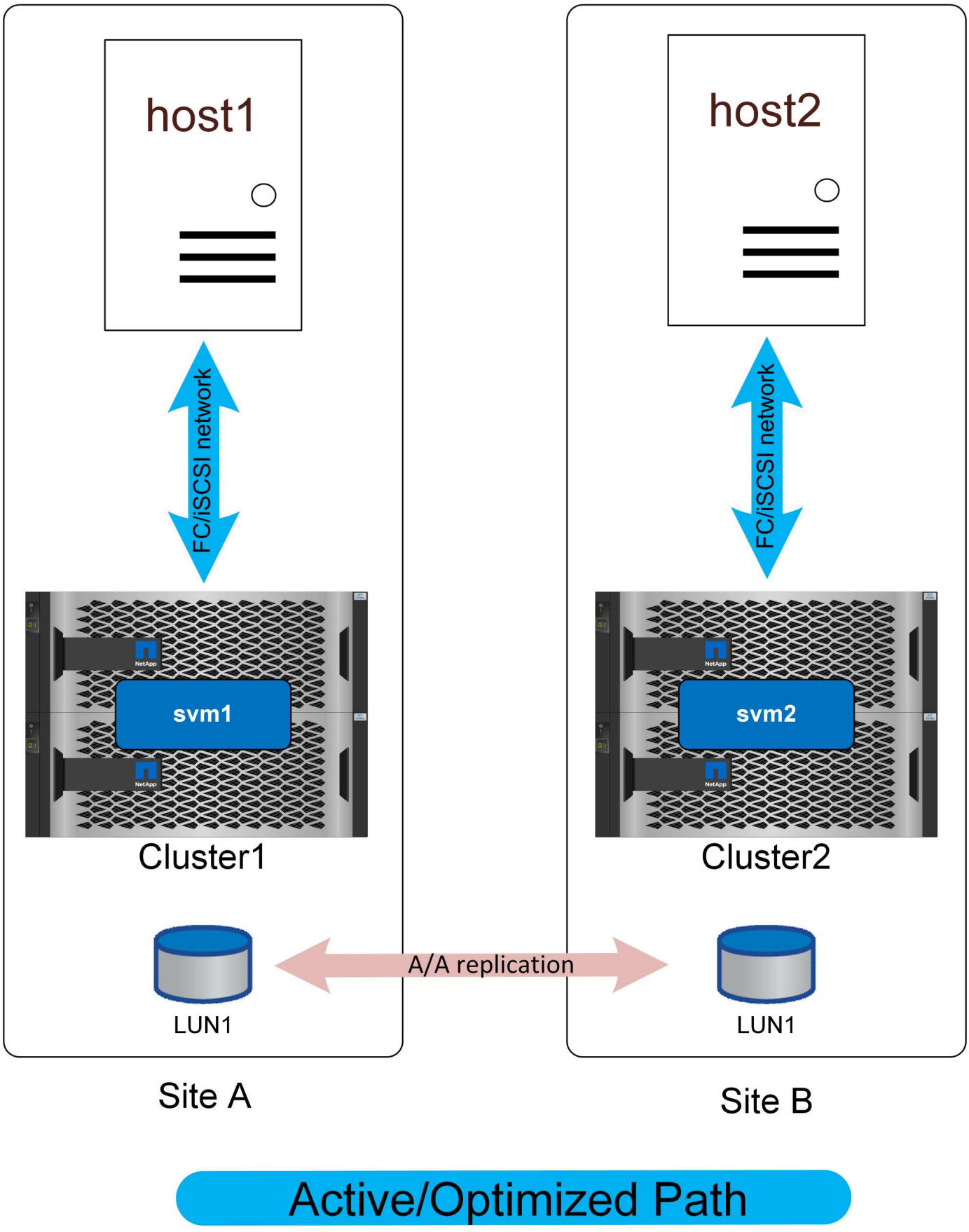
Uma configuração ASA com acesso não uniforme funcionaria em grande parte da mesma forma que faria com o AFF. Com acesso uniforme, o IO estaria atravessando a WAN. Isto pode ou não ser desejável.

Se os dois locais estivessem a 100 metros de distância com conetividade de fibra, não deveria haver latência adicional detetável cruzando a WAN, mas se os locais estivessem a uma distância longa, então o desempenho de leitura sofreria em ambos os locais. Em contraste, com o AFF, esses caminhos de cruzamento de WAN só seriam usados se não houvesse caminhos locais disponíveis e o desempenho do dia-a-dia seria melhor porque todo o IO seria IO local. O ASA com rede de acesso não uniforme seria uma opção para obter os benefícios de custo e recursos do ASA sem incorrer em uma penalidade de acesso à latência entre locais.

O ASA com SM-as em uma configuração de baixa latência oferece dois benefícios interessantes. Primeiro, ele basicamente **dobra** a performance de qualquer host porque a e/S pode ser atendida por duas vezes mais controladoras usando o dobro de caminhos. Em segundo lugar, em um ambiente de local único, ele oferece disponibilidade extrema porque todo um sistema de storage pode ser perdido sem interromper o acesso de host.

Acesso não uniforme

A rede de acesso não uniforme significa que cada host só tem acesso às portas no sistema de storage local. A SAN não é estendida entre sites (ou domínios de falha dentro do mesmo site).



O principal benefício dessa abordagem é a simplicidade da SAN - você elimina a necessidade de estender uma SAN pela rede. Alguns clientes não têm conetividade de baixa latência suficiente entre locais ou não têm

infraestrutura para túnel do tráfego SAN FC em uma rede entre locais.

A desvantagem para o acesso não uniforme é que certos cenários de falha, incluindo a perda do link de replicação, resultarão em alguns hosts perdendo acesso ao armazenamento. Os aplicativos que são executados como instâncias únicas, como um banco de dados não agrupado, que inerentemente está sendo executado apenas em um único host em qualquer montagem, falharão se a conetividade de armazenamento local for perdida. Os dados ainda seriam protegidos, mas o servidor de banco de dados não teria mais acesso. Ele precisaria ser reiniciado em um local remoto, de preferência através de um processo automatizado. Por exemplo, o VMware HA pode detetar uma situação de todos os caminhos em um servidor e reiniciar uma VM em outro servidor onde os caminhos estão disponíveis.

Em contraste, um aplicativo em cluster, como o Oracle RAC, pode fornecer um serviço que está disponível simultaneamente em dois locais diferentes. Perder um site não significa perda do serviço do aplicativo como um todo. As instâncias ainda estão disponíveis e em execução no local sobrevivente.

Em muitos casos, a sobrecarga de latência adicional de um aplicativo que acessa o storage em um link local a local seria inaceitável. Isso significa que a disponibilidade aprimorada de redes uniformes é mínima, uma vez que a perda de armazenamento em um local levaria à necessidade de encerrar serviços nesse local com falha de qualquer maneira.

> Há caminhos redundantes pelo cluster local que não são mostrados nesses diagramas por uma questão de simplicidade. Os sistemas de storage da ONTAP estão HA, portanto, uma falha da controladora não deve resultar em falha do local. Deve apenas resultar em uma mudança na qual os caminhos locais são usados no site afetado.

Configurações Oracle

Visão geral

O uso da sincronização ativa do SnapMirror não necessariamente adiciona ou altera quaisquer práticas recomendadas para operar um banco de dados.

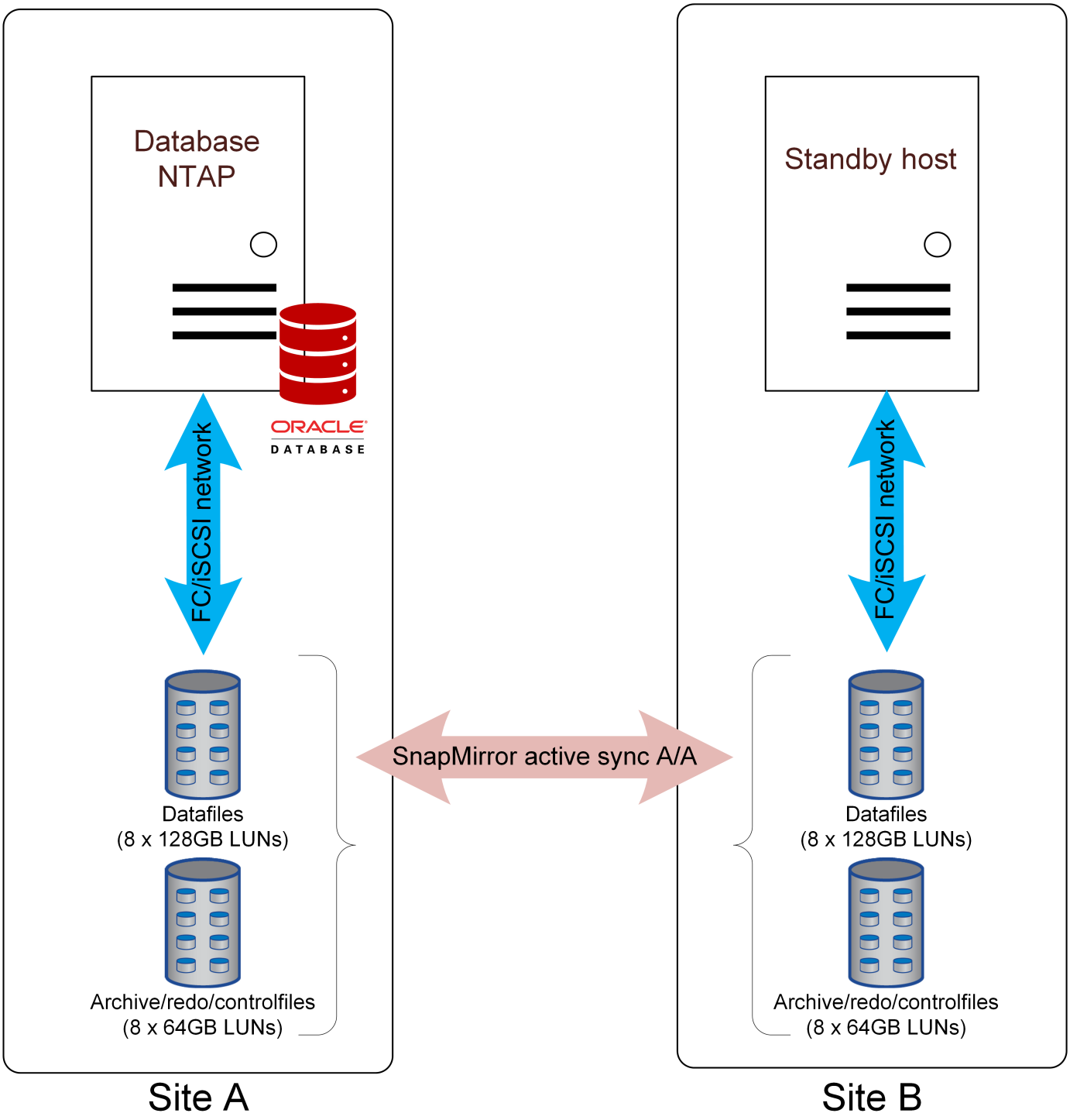
A melhor arquitetura depende dos requisitos de negócios. Por exemplo, se o objetivo é ter proteção RPO igual a 0 contra a perda de dados, mas o rto estiver relaxado, o uso de bancos de dados de Instância única Oracle e a replicação dos LUNs com SM-as pode ser suficiente e menos caro de um padrão de licenciamento Oracle. A falha do local remoto não interromperia as operações e a perda do local principal resultaria em LUNs no local sobrevivente que estão on-line e prontos para serem usados.

Se o rto fosse mais rigoroso, a automação ativo-passivo básica por meio de scripts ou clusterware, como pacemaker ou Ansible, melhoraria o tempo de failover. Por exemplo, o VMware HA pode ser configurado para detetar falha de VM no local principal e ativar a VM no local remoto.

Finalmente, para um failover extremamente rápido, o Oracle RAC poderia ser implantado em todos os locais. O rto seria essencialmente zero porque o banco de dados estaria online e disponível em ambos os sites em todos os momentos.

Instância única Oracle

Os exemplos explicados abaixo mostram algumas das muitas opções para implantar bancos de dados de Instância única Oracle com replicação de sincronização ativa do SnapMirror.



Failover com um SO pré-configurado

O SnapMirror active Sync fornece uma cópia síncrona dos dados no local de recuperação de desastres, 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 pacemaker, costumam ser 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) colocará os bancos de dados on-line no site alternativo. Uma opção é criar servidores em espera pré-configurados para os recursos 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. Detectar falha do local principal
2. Realize a descoberta de LUNs FC ou iSCSI
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.

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. Detectar falha do local principal
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 locais. Em caso de desastre, as máquinas virtuais podem ser colocadas on-line no local de recuperação de desastres após o switchover.

Proteção contra falha de storage

O diagrama acima mostra o uso "acesso não uniforme"do , em que a SAN não é estendida nos locais. Isso pode ser mais simples de configurar e, em alguns casos, pode ser a única opção, dada a capacidade de SAN atual, mas também significa que a falha do sistema de storage primário causaria uma interrupção do banco de dados até que o aplicativo fosse failover.

Para obter resiliência adicional, a solução poderia ser implantada com "acesso uniforme"o . Isso permitiria que os aplicativos continuassem operando usando os caminhos anunciados a partir do site oposto.

Oracle Extended RAC

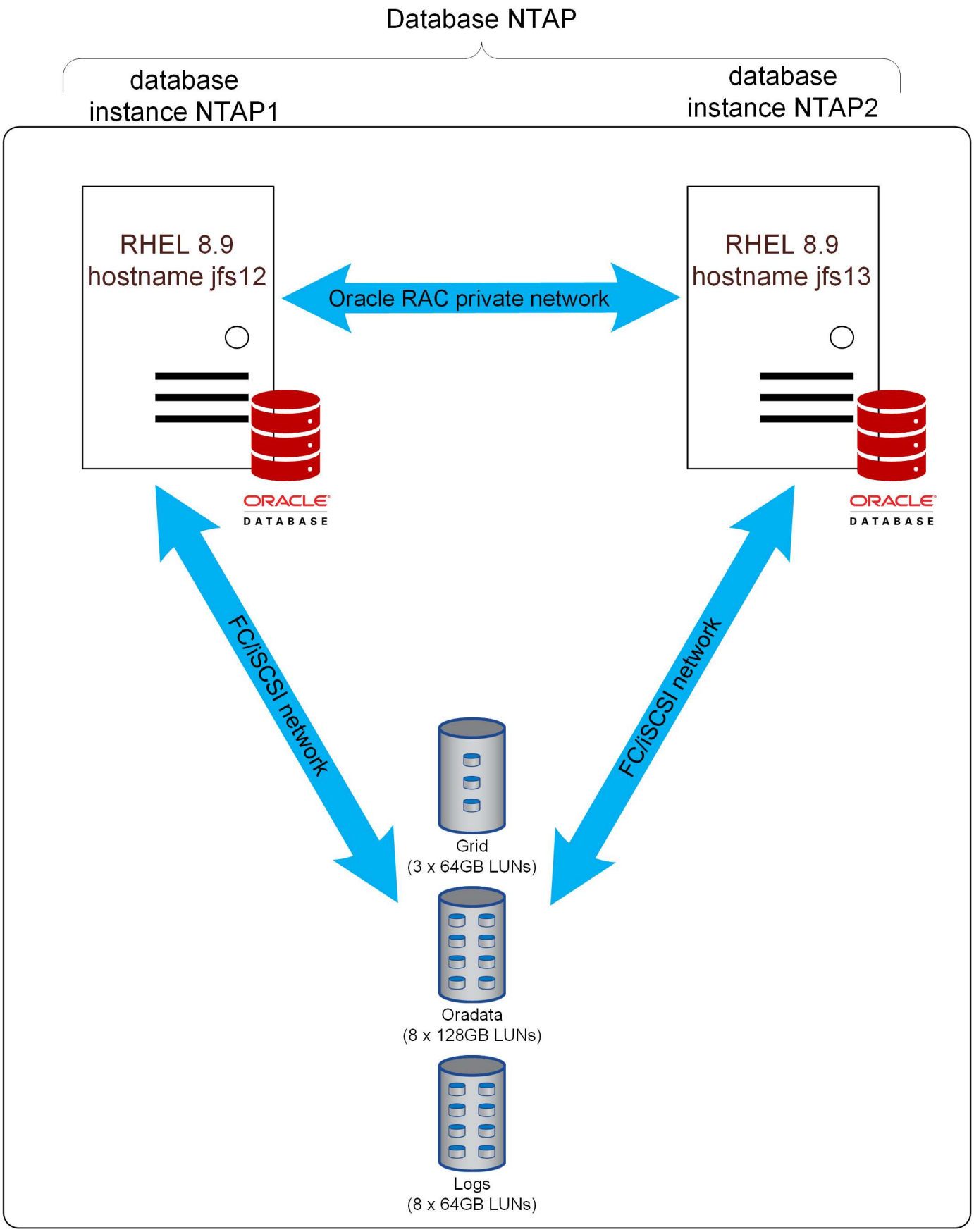
Muitos clientes otimizam seu rto alongando um cluster do Oracle RAC entre locais, gerando uma configuração totalmente ativo-ativo. O projeto geral se torna mais complicado porque deve incluir o gerenciamento de quórum do Oracle RAC.

O RAC estendido tradicional em cluster contou com o espelhamento ASM para fornecer proteção de dados. Essa abordagem funciona, mas também requer muitas etapas de configuração manual e impõe sobrecarga na infraestrutura de rede. Em contraste, permitir que o SnapMirror ative Sync assuma a responsabilidade pela replicação de dados simplifica significativamente a solução. Operações como sincronização, ressincronização

após interrupções, failovers e gerenciamento de quórum são mais fáceis, e a SAN não precisa ser distribuída entre locais, o que simplifica o design e o gerenciamento da SAN.

Replicação

A chave para entender a funcionalidade RAC no SnapMirror ative Sync é visualizar o armazenamento como um único conjunto de LUNs hospedados no armazenamento espelhado. Por exemplo:



Não há cópia primária ou cópia espelhada. Logicamente, há apenas uma única cópia de cada LUN e esse LUN está disponível em caminhos SAN localizados em dois sistemas de armazenamento diferentes. Do ponto de vista do host, não há failovers de storage; em vez disso, há alterações de caminho. Vários eventos de falha

podem levar à perda de certos caminhos para o LUN, enquanto outros caminhos permanecem on-line. A sincronização ativa do SnapMirror garante que os mesmos dados estejam disponíveis em todos os caminhos operacionais.

Configuração de armazenamento

Neste exemplo de configuração, os discos ASM são configurados da mesma forma que seriam em qualquer configuração RAC de local único no storage empresarial. Como o sistema de armazenamento fornece proteção de dados, a redundância externa ASM seria usada.

Uniforme vs acesso não ininformado

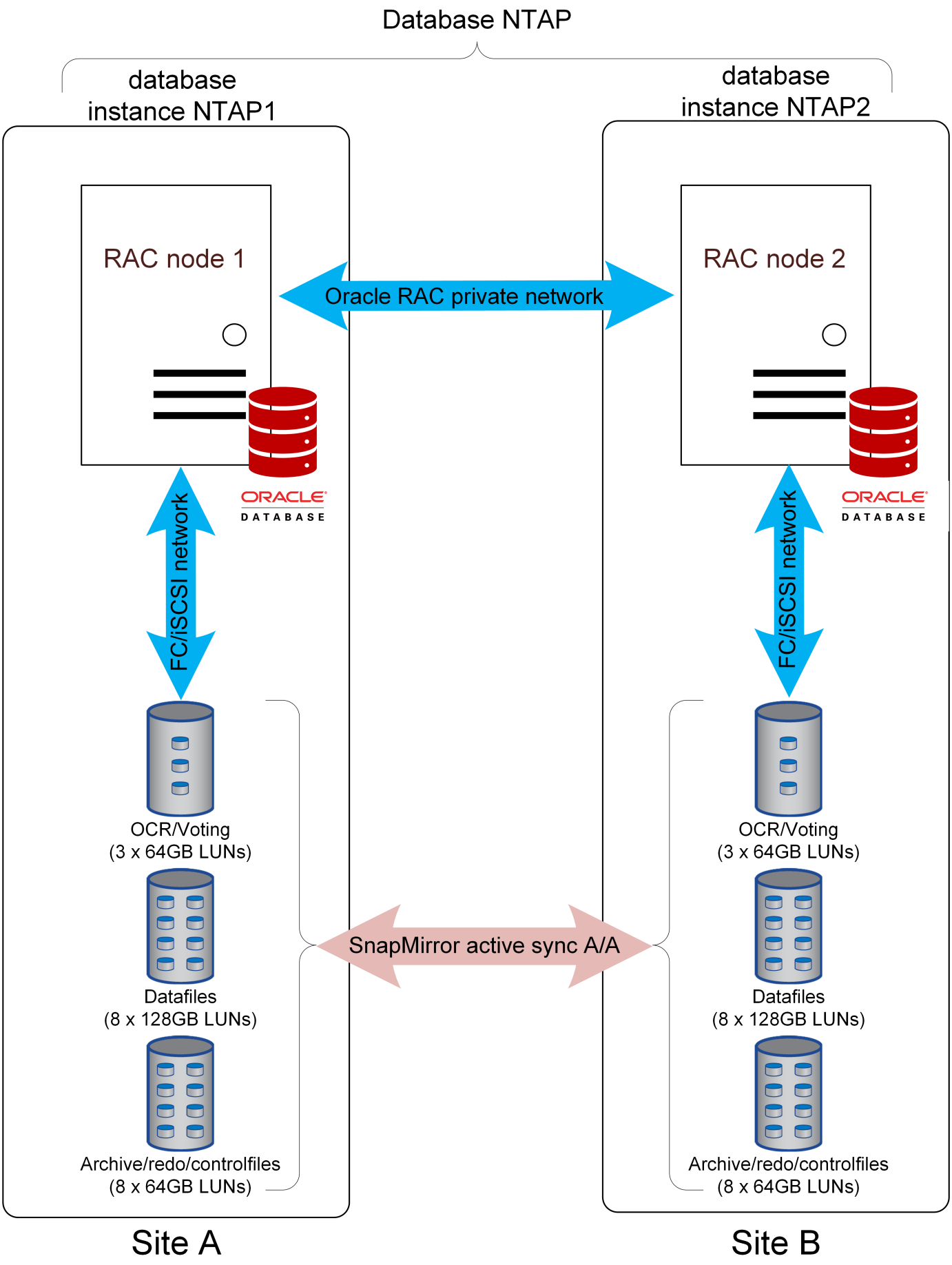
A consideração mais importante com o Oracle RAC na sincronização ativa do SnapMirror é usar acesso uniforme ou não uniforme.

O acesso uniforme significa que cada host pode ver caminhos em ambos os clusters. O acesso não uniforme significa que os hosts só podem ver caminhos para o cluster local.

Nenhuma das opções é especificamente recomendada ou desencorajada. Alguns clientes têm fibra escura prontamente disponível para conetar sites, outros não têm essa conetividade ou sua infraestrutura de SAN não oferece suporte a um ISL de longa distância.

Acesso não uniforme

O acesso não uniforme é mais simples de configurar do ponto de vista da SAN.


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
ORACLE DATABASE
ORACLE DATABASE
FC/iSCSI network
FC/iSCSI network
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B

A principal desvantagem "acesso não uniforme"da abordagem é que a perda de conetividade ONTAP site a site ou a perda de um sistema de storage resultará na perda de instâncias de banco de dados em um local. Isso obviamente não é desejável, mas pode ser um risco aceitável em troca de uma configuração SAN mais simples.

Acesso uniforme

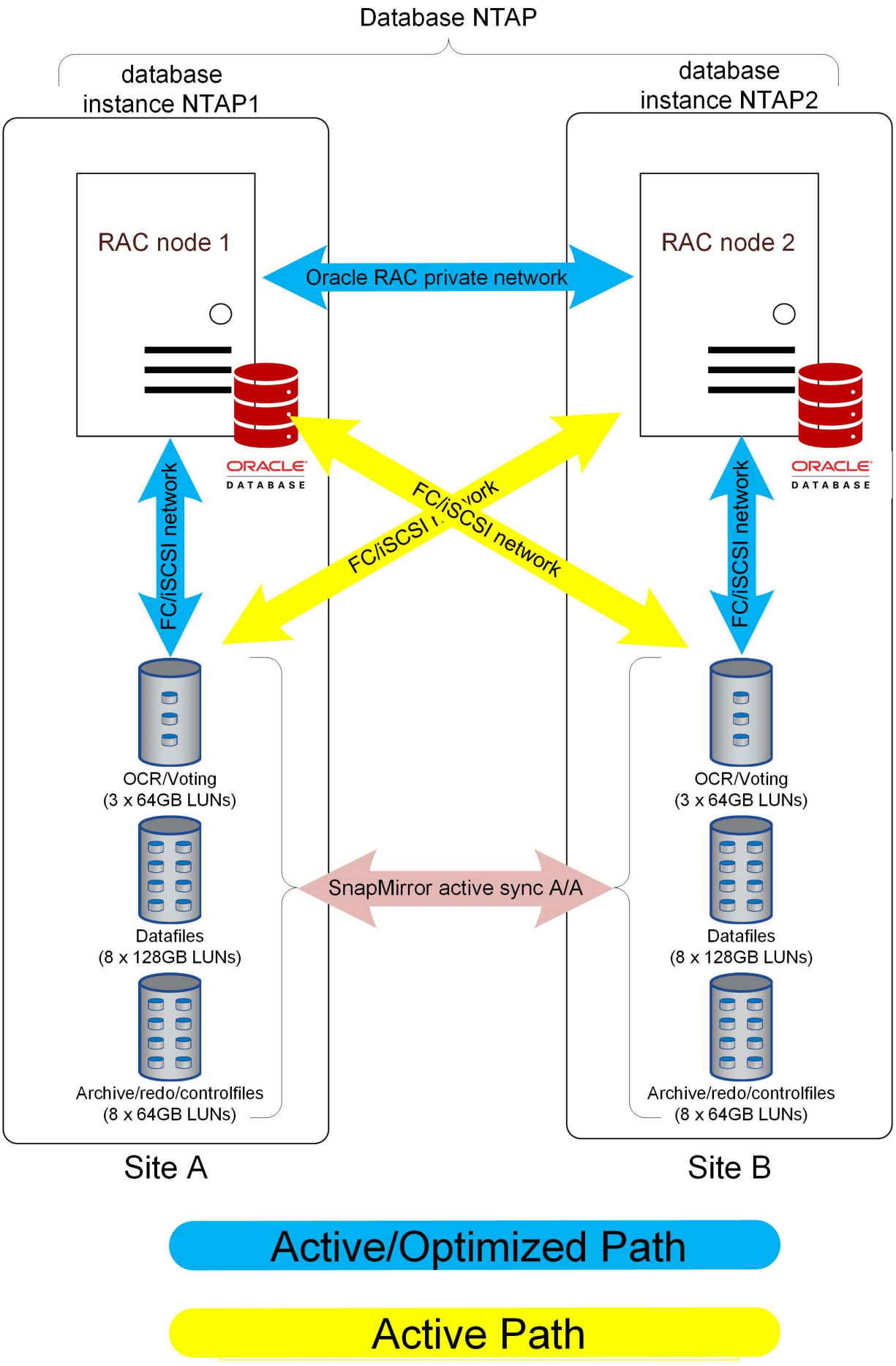
O acesso uniforme requer a extensão da SAN entre os locais. O principal benefício é que a perda de um sistema de storage não resultará na perda de uma instância de banco de dados. Em vez disso, isso resultaria em uma mudança multipathing em que os caminhos estão atualmente em uso.

Existem várias maneiras de configurar o acesso não uniforme.

> Nos diagramas abaixo, há também caminhos ativos, mas não otimizados, que seriam usados durante falhas simples do controlador, mas esses caminhos não são mostrados no interesse de simplificar os diagramas.

AFF com definições de proximidade

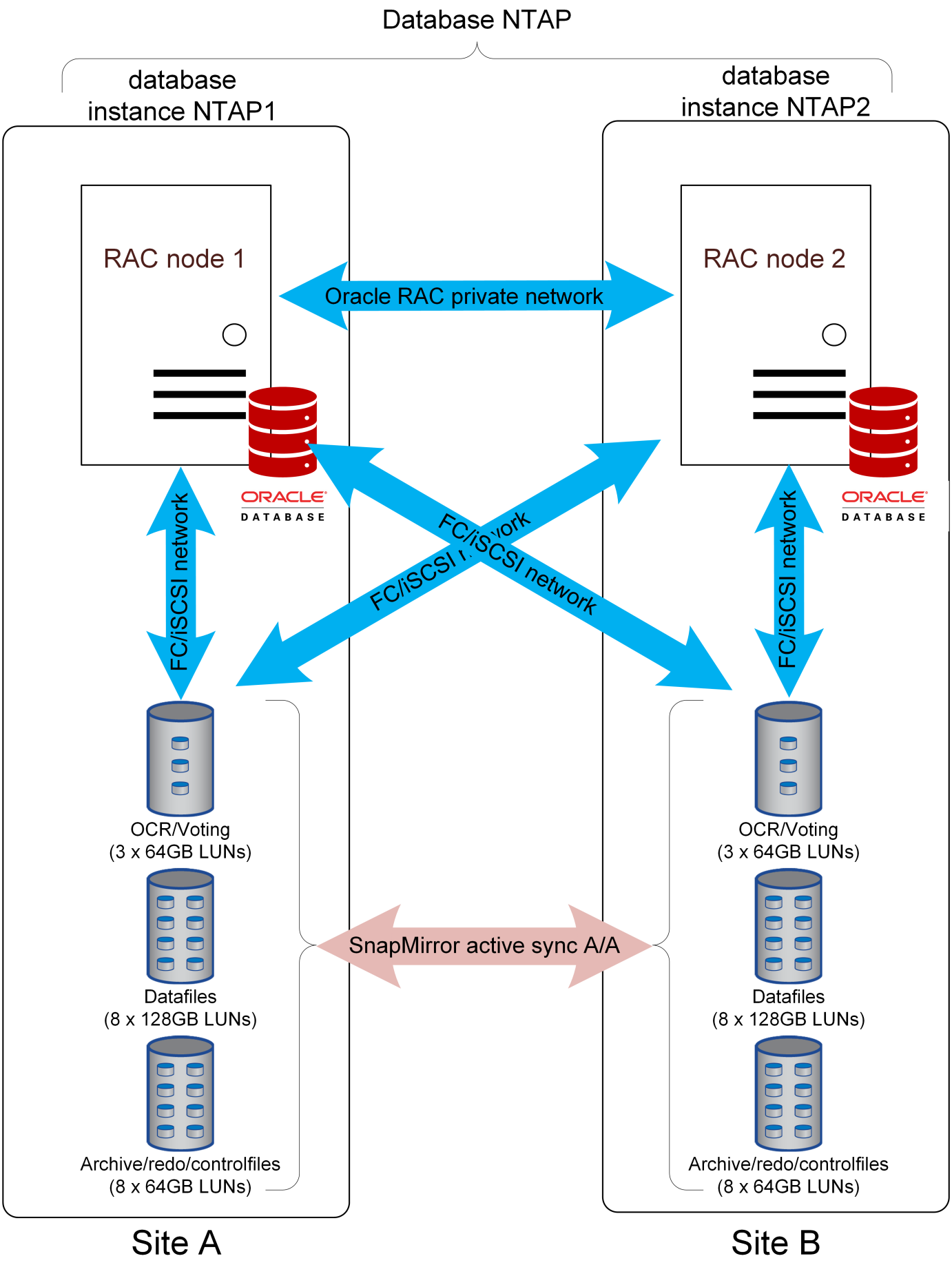
Se houver latência significativa entre sites, os sistemas AFF podem ser configurados com configurações de proximidade do host. Isso permite que cada sistema de armazenamento esteja ciente de quais hosts são locais e quais são remotos e atribua prioridades de caminho adequadamente.


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
ORACLE DATABASE
ORACLE DATABASE
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B
Active/Optimized Path
Active Path

Na operação normal, cada instância do Oracle usaria preferencialmente os caminhos ativos/otimizados locais. O resultado é que todas as leituras seriam atendidas pela cópia local dos blocos. Isso produz a menor latência possível. O write IO é enviado de forma semelhante para o controlador local. O IO ainda deve ser replicado antes de ser reconhecido e, portanto, ainda incorreria na latência adicional de cruzar a rede local a local, mas isso não pode ser evitado em uma solução de replicação síncrona.

ASA / AFF sem definições de proximidade

Se não houver latência significativa entre sites, os sistemas AFF podem ser configurados sem configurações de proximidade do host ou ASA podem ser usados.


Database NTAP
database instance NTAP1
database instance NTAP2
RAC node 1
RAC node 2
Oracle RAC private network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
FC/iSCSI network
ORACLE DATABASE
ORACLE DATABASE
OCR/Voting (3 x 64GB LUNs)
OCR/Voting (3 x 64GB LUNs)
SnapMirror active sync A/A
Datafiles (8 x 128GB LUNs)
Datafiles (8 x 128GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Archive/redo/controlfiles (8 x 64GB LUNs)
Site A
Site B

Cada host poderá usar todos os caminhos operacionais em ambos os sistemas de storage. Isso potencialmente melhora o desempenho de maneira significativa, permitindo que cada host aproveite o potencial de desempenho de dois clusters, e não apenas um.

Com o ASA, não só todos os caminhos para ambos os clusters seriam considerados ativos e otimizados, como também os caminhos nos controladores do parceiro estariam ativos. O resultado seria caminhos SAN all-ativos em todo o cluster, o tempo todo.

> Os sistemas ASA também podem ser usados em uma configuração de acesso não uniforme. Uma vez que não existem caminhos entre locais, não haveria impactos no desempenho resultante do IO cruzando o ISL.

Desempate do RAC

Embora o RAC estendido usando o SnapMirror ative Sync seja uma arquitetura simétrica com relação ao IO, há uma exceção que está conetada ao gerenciamento de split-brain.

O que acontece se o link de replicação for perdido e nenhum dos sites tiver quórum? O que deve acontecer? Esta pergunta se aplica ao comportamento do Oracle RAC e do ONTAP. Se as alterações não puderem ser replicadas nos sites e você quiser retomar as operações, um dos sites terá que sobreviver e o outro site terá que ficar indisponível.

O "ONTAP Mediador" atende a esse requisito na camada ONTAP. Existem várias opções para quebra de binário RAC.

Tiebreakers Oracle

O melhor método para gerenciar os riscos do Oracle RAC dividido é usar um número ímpar de nós RAC, de preferência pelo uso de um desempate de site 3rd. Se um site 3rd não estiver disponível, a instância tiebreaker pode ser colocada em um local dos dois sites, designando-o efetivamente um local sobrevivente preferido.

Oracle e CSS_critical

Com um número par de nós, o comportamento padrão do Oracle RAC é que um dos nós no cluster será considerado mais importante do que os outros nós. O local com esse nó de maior prioridade sobreviverá ao isolamento do local, enquanto os nós do outro local serão despejados. A priorização é baseada em vários fatores, mas você também pode controlar esse comportamento usando a `css_critical` configuração.

Na "exemplo" arquitetura, os nomes de host para os nós RAC são jfs12 e jfs13. As definições atuais para `css_critical` são as seguintes:

```
[root@jfs12 ~]# /grid/bin/crsctl get server css_critical
CRS-5092: Current value of the server attribute CSS_CRITICAL is no.

[root@jfs13 trace]# /grid/bin/crsctl get server css_critical
CRS-5092: Current value of the server attribute CSS_CRITICAL is no.
```

Se você quiser que o site com jfs12 seja o site preferido, altere esse valor para sim em um nó De site e reinicie os serviços.

```
[root@jfs12 ~]# /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.

[root@jfs12 ~]# /grid/bin/crsctl stop crs
CRS-2791: Starting shutdown of Oracle High Availability Services-managed
resources on 'jfs12'
CRS-2673: Attempting to stop 'ora.crsd' on 'jfs12'
CRS-2790: Starting shutdown of Cluster Ready Services-managed resources on
server 'jfs12'
CRS-2673: Attempting to stop 'ora.ntap.ntappdb1.pdb' on 'jfs12'
…
CRS-2673: Attempting to stop 'ora.gipcd' on 'jfs12'
CRS-2677: Stop of 'ora.gipcd' on 'jfs12' succeeded
CRS-2793: Shutdown of Oracle High Availability Services-managed resources
on 'jfs12' has completed
CRS-4133: Oracle High Availability Services has been stopped.

[root@jfs12 ~]# /grid/bin/crsctl start crs
CRS-4123: Oracle High Availability Services has been started.
```

Cenários de falha

Visão geral

Planejar uma arquitetura completa de aplicativos de sincronização ativa do SnapMirror requer entender como o SM-as responderá em vários cenários de failover planejados e não planejados.

Para os exemplos a seguir, suponha que o site A esteja configurado como o site preferido.

Perda de conetividade de replicação

Se a replicação SM-as for interrompida, não é possível concluir a e/S de gravação porque seria impossível que um cluster replique alterações no local oposto.

Local A (local preferido)

O resultado da falha do link de replicação no site preferido será uma pausa de aproximadamente 15 segundos no processamento de e/S de gravação, à medida que o ONTAP tenta novamente as operações de gravação replicadas antes de determinar que o link de replicação é genuinamente inacessível. Após os 15 segundos decorridos, o Site Um sistema retoma o processamento de e/S de leitura e escrita. Os caminhos de SAN não serão alterados e os LUNs permanecerão online.

Local B

Como o local B não é o site preferido de sincronização ativa do SnapMirror, seus caminhos de LUN ficarão indisponíveis após cerca de 15 segundos.

Falha do sistema de storage

O resultado de uma falha do sistema de armazenamento é quase idêntico ao resultado da perda do link de replicação. O local sobrevivente deve experimentar uma pausa de IO de aproximadamente 15 segundos. Uma vez decorrido esse período de 15 segundos, o IO será retomado nesse site como de costume.

Perda do mediador

O serviço mediador não controla diretamente as operações de storage. Ele funciona como um caminho de controle alternativo entre clusters. Ele existe principalmente para automatizar o failover sem o risco de um cenário de divisão cerebral. Em operação normal, cada cluster replica alterações em seu parceiro, e cada cluster pode verificar se o cluster do parceiro está on-line e fornecendo dados. Se o link de replicação falhar, a replicação cessaria.

O motivo pelo qual um mediador é necessário para o failover automatizado seguro é porque, de outra forma, seria impossível que um cluster de storage pudesse determinar se a perda de comunicação bidirecional foi o resultado de uma interrupção da rede ou falha real do storage.

O mediador fornece um caminho alternativo para cada cluster para verificar a integridade de seu parceiro. Os cenários são os seguintes:

- Se um cluster puder entrar em Contato diretamente com seu parceiro, os serviços de replicação estarão operacionais. Nenhuma ação necessária.
- Se um site preferido não puder entrar em Contato diretamente com seu parceiro ou por meio do mediador, ele assumirá que ele está realmente indisponível ou foi isolado e que levou seus caminhos LUN off-line. O site preferido continuará lançando o estado RPO/0 e continuará processando e/S de leitura e gravação.
- Se um site não preferencial não puder entrar em Contato diretamente com seu parceiro, mas puder contatá-lo por meio do mediador, ele tomará seus caminhos off-line e aguardará o retorno da conexão de replicação.
- Se um site não preferencial não puder entrar em Contato com seu parceiro diretamente ou por meio de um mediador operacional, ele assumirá que o parceiro está realmente indisponível ou foi isolado e que tomou seus caminhos LUN off-line. O site não preferencial lançará o estado RPO 0 e continuará processando e/S de leitura e gravação. Ele assumirá o papel da fonte de replicação e se tornará o novo site preferido.

Se o mediador não estiver totalmente disponível:

- A falha dos serviços de replicação por qualquer motivo, incluindo a falha do sistema de storage ou local não preferido, resultará no lançamento do estado RPO/0 e no reinício do processamento de e/S de leitura e gravação. O site não preferencial tomará seus caminhos off-line.
- A falha do site preferido resultará em uma falha porque o site não-preferido não será capaz de verificar se o site oposto está realmente off-line e, portanto, não seria seguro para o site não-preferido retomar os serviços.

Restauração de serviços

Depois que uma falha é resolvida, como restaurar a conetividade site-a-site ou ligar um sistema com falha, os pontos de extremidade de sincronização ativa do SnapMirror detetarão automaticamente a presença de uma relação de replicação com defeito e o devolverão ao estado RPO-0. Uma vez que a replicação síncrona for restabelecida, os caminhos com falha ficarão online novamente.

Em muitos casos, os aplicativos em cluster detetarão automaticamente o retorno de caminhos com falha, e esses aplicativos também voltarão online. Em outros casos, pode ser necessária uma análise SAN no nível do

host ou os aplicativos podem precisar ser colocados online manualmente. Depende do aplicativo e como ele é configurado e, em geral, essas tarefas podem ser facilmente automatizadas. O próprio ONTAP é com autorrecuperação e não deve exigir a intervenção do usuário para retomar as operações de storage RPO de 0.

Failover manual

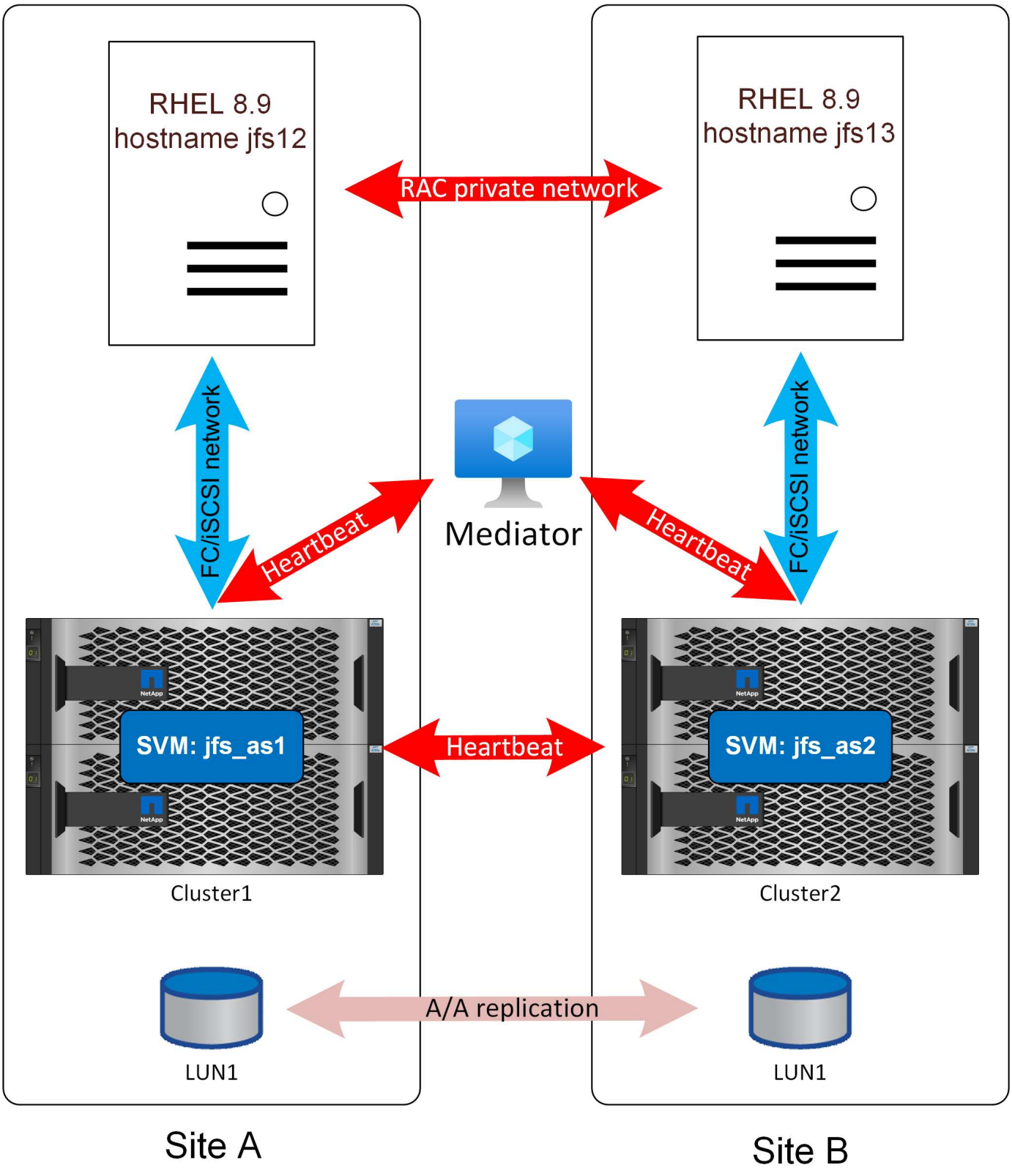
Alterar o local preferido requer uma operação simples. A e/S pausa por um segundo ou dois como autoridade sobre os switches de comportamento de replicação entre clusters, mas a e/S não é afetada.

Arquitetura de amostra

Os exemplos de falha detalhados mostrados nestas seções são baseados na arquitetura mostrada abaixo.

> Esta é apenas uma das muitas opções para bancos de dados Oracle na sincronização ativa do SnapMirror. Este design foi escolhido porque ilustra alguns dos cenários mais complicados.

Neste design, suponha que o local A esteja definido no "site preferido".



Falha na interconexão RAC

A perda do link de replicação do Oracle RAC produzirá um resultado semelhante à perda de conetividade do SnapMirror, exceto que os tempos limite serão menores por padrão. Sob as configurações padrão, um nó Oracle RAC aguardará 200 segundos após a perda

de conetividade de armazenamento antes de despejar, mas só aguardará 30 segundos após a perda do batimento cardíaco da rede RAC.

As mensagens CRS são semelhantes às mostradas abaixo. Você pode ver o intervalo de tempo limite de 30 segundos. Uma vez que CSS_critical foi definido em jfs12, localizado no local A, que será o local para sobreviver e jfs13 no local B será despejado.

```
2024-09-12 10:56:44.047 [ONMD(3528)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 6.980 seconds
2024-09-12 10:56:48.048 [ONMD(3528)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 2.980 seconds
2024-09-12 10:56:51.031 [ONMD(3528)]CRS-1607: Node jfs13 is being evicted
in cluster incarnation 621599354; details at (:CSSNM00007:) in
/gridbase/diag/crs/jfs12/crs/trace/onmd.trc.
2024-09-12 10:56:52.390 [CRSD(6668)]CRS-7503: The Oracle Grid
Infrastructure process 'crsd' observed communication issues between node
'jfs12' and node 'jfs13', interface list of local node 'jfs12' is
'192.168.30.1:33194;', interface list of remote node 'jfs13' is
'192.168.30.2:33621;'.
2024-09-12 10:56:55.683 [ONMD(3528)]CRS-1601: CSSD Reconfiguration
complete. Active nodes are jfs12 .
2024-09-12 10:56:55.722 [CRSD(6668)]CRS-5504: Node down event reported for
node 'jfs13'.
2024-09-12 10:56:57.222 [CRSD(6668)]CRS-2773: Server 'jfs13' has been
removed from pool 'Generic'.
2024-09-12 10:56:57.224 [CRSD(6668)]CRS-2773: Server 'jfs13' has been
removed from pool 'ora.NTAP'.
```

Falha de comunicação do SnapMirror

Se o link de replicação de sincronização ativa do SnapMirror, a e/S de gravação não puder ser concluída porque seria impossível que um cluster replique alterações no local oposto.

Local A

O resultado de uma falha no link de replicação no local A será uma pausa de aproximadamente 15 segundos no processamento de e/S de gravação, já que o ONTAP tenta replicar gravações antes de determinar que o link de replicação está genuinamente inoperável. Após os 15 segundos decorridos, o cluster do ONTAP no local A retoma o processamento de e/S de leitura e gravação. Os caminhos de SAN não serão alterados e os LUNs permanecerão online.

Local B

Como o local B não é o site preferido de sincronização ativa do SnapMirror, seus caminhos de LUN ficarão indisponíveis após cerca de 15 segundos.

O link de replicação foi cortado no carimbo de data/hora 15:19:44. O primeiro aviso do Oracle RAC chega 100 segundos depois, quando o tempo limite de 200 segundos (controlado pelo parâmetro Oracle RAC disktimeout) se aproxima.

```
2024-09-10 15:21:24.702 [ONMD(2792)]CRS-1615: No I/O has completed after
50% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 99340 milliseconds.
2024-09-10 15:22:14.706 [ONMD(2792)]CRS-1614: No I/O has completed after
75% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 49330 milliseconds.
2024-09-10 15:22:44.708 [ONMD(2792)]CRS-1613: No I/O has completed after
90% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 19330 milliseconds.
2024-09-10 15:23:04.710 [ONMD(2792)]CRS-1604: CSSD voting file is offline:
/dev/mapper/grid2; details at (:CSSNM00058:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc.
2024-09-10 15:23:04.710 [ONMD(2792)]CRS-1606: The number of voting files
available, 0, is less than the minimum number of voting files required, 1,
resulting in CSSD termination to ensure data integrity; details at
(:CSSNM00018:) in /gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-10 15:23:04.716 [ONMD(2792)]CRS-1699: The CSS daemon is
terminating due to a fatal error from thread:
clssnmvDiskPingMonitorThread; Details at (:CSSSC00012:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-10 15:23:04.731 [OCSSD(2794)]CRS-1652: Starting clean up of CRSD
resources.
```

Assim que o tempo limite do disco de votação de 200 segundos for atingido, esse nó Oracle RAC se despejará do cluster e reiniciará.

Falha total de interconetividade de rede

Se o link de replicação entre sites for completamente perdido, a sincronização ativa do SnapMirror e a conetividade do Oracle RAC serão interrompidas.

A deteção de split-brain do Oracle RAC tem uma dependência do batimento cardíaco do armazenamento do Oracle RAC. Se a perda de conetividade local a local resultar na perda simultânea dos serviços de replicação de armazenamento e batimento cardíaco da rede RAC, o resultado será que os sites RAC não poderão se comunicar entre locais através da interconexão RAC ou dos discos de votação RAC. O resultado em um conjunto de nós com dormência uniforme pode ser despejo de ambos os sites sob configurações padrão. O comportamento exato dependerá da sequência de eventos e do tempo das pesquisas de batimento cardíaco da rede RAC e do disco.

O risco de uma interrupção no 2 local pode ser resolvido de duas maneiras. Primeiro, uma "desempate" configuração pode ser usada.

Se um site 3rd não estiver disponível, esse risco pode ser resolvido ajustando o parâmetro misscount no cluster RAC. Sob os padrões, o tempo limite do batimento cardíaco da rede RAC é de 30 segundos. Isso normalmente é usado pelo RAC para identificar os nós RAC com falha e removê-los do cluster. Ele também

tem uma conexão com o heartbeat do disco de votação.

Se, por exemplo, o canal que transporta tráfego intersite para Oracle RAC e serviços de replicação de armazenamento for cortado por uma retroescavadeira, a contagem regressiva de 30 segundos de misscount começará. Se o nó do local preferencial RAC não puder restabelecer o Contato com o local oposto dentro de 30 segundos, e ele também não puder usar os discos de votação para confirmar que o local oposto está para baixo dentro dessa mesma janela de 30 segundos, os nós do local preferido também serão despejados. O resultado é uma interrupção completa do banco de dados.

Dependendo de quando a polling do misscount ocorre, 30 segundos podem não ser tempo suficiente para que a sincronização ativa do SnapMirror expire e permita que o armazenamento no site preferido retome os serviços antes que a janela de 30 segundos expire. Esta janela de 30 segundos pode ser aumentada.

```
[root@jfs12 ~]# /grid/bin/crsctl set css misscount 100
CRS-4684: Successful set of parameter misscount to 100 for Cluster
Synchronization Services.
```

Esse valor permite que o sistema de armazenamento no site preferido retome as operações antes que o tempo limite de contagem de erros expire. O resultado será, então, despejo apenas dos nós no local onde os caminhos LUN foram removidos. Exemplo abaixo:

```
2024-09-12 09:50:59.352 [ONMD(681360)]CRS-1612: Network communication with
node jfs13 (2) has been missing for 50% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 49.570 seconds
2024-09-12 09:51:10.082 [CRSD(682669)]CRS-7503: The Oracle Grid
Infrastructure process 'crsd' observed communication issues between node
'jfs12' and node 'jfs13', interface list of local node 'jfs12' is
'192.168.30.1:46039;', interface list of remote node 'jfs13' is
'192.168.30.2:42037;'.
2024-09-12 09:51:24.356 [ONMD(681360)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 24.560 seconds
2024-09-12 09:51:39.359 [ONMD(681360)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 9.560 seconds
2024-09-12 09:51:47.527 [OHASD(680884)]CRS-8011: reboot advisory message
from host: jfs13, component: cssagent, with time stamp: L-2024-09-12-
09:51:47.451
2024-09-12 09:51:47.527 [OHASD(680884)]CRS-8013: reboot advisory message
text: oracssdagent is about to reboot this node due to unknown reason as
it did not receive local heartbeats for 10470 ms amount of time
2024-09-12 09:51:48.925 [ONMD(681360)]CRS-1632: Node jfs13 is being
removed from the cluster in cluster incarnation 621596607
```

O Oracle Support desencoraja fortemente a alteração com os parâmetros misscount ou disktimeout para resolver problemas de configuração. A alteração desses parâmetros pode, no entanto, ser garantida e inevitável em muitos casos, incluindo configurações de inicialização por SAN, virtualizadas e replicação de

armazenamento. Se, por exemplo, você tiver problemas de estabilidade com uma rede SAN ou IP que resultasse em despejos RAC, você deve corrigir o problema subjacente e não cobrar os valores do misscount ou disktimeout. Alterar tempos limite para resolver erros de configuração está mascarando um problema, não resolvendo um problema. Alterar esses parâmetros para configurar adequadamente um ambiente RAC com base em aspetos de design da infraestrutura subjacente é diferente e é consistente com as declarações de suporte Oracle. Com a inicialização SAN, é comum ajustar o misscount até 200 para corresponder ao disktimeout. "este link"Consulte para obter informações adicionais.

Falha do local

O resultado de uma falha do sistema de armazenamento ou do local é quase idêntico ao resultado da perda do link de replicação. O site sobrevivente deve experimentar uma pausa de IO de cerca de 15 segundos nas gravações. Uma vez decorrido esse período de 15 segundos, o IO será retomado nesse site como de costume.

Se apenas o sistema de armazenamento tiver sido afetado, o nó Oracle RAC no site com falha perderá os serviços de armazenamento e entrará na mesma contagem regressiva de tempo limite de disco de 200 segundos antes da remoção e reinicialização subsequente.

```
2024-09-11 13:44:38.613 [ONMD(3629)]CRS-1615: No I/O has completed after
50% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 99750 milliseconds.
2024-09-11 13:44:51.202 [ORAAGENT(5437)]CRS-5011: Check of resource "NTAP"
failed: details at "(:CLSN00007:)" in
"/gridbase/diag/crs/jfs13/crs/trace/crsd_oraagent_oracle.trc"
2024-09-11 13:44:51.798 [ORAAGENT(75914)]CRS-8500: Oracle Clusterware
ORAAGENT process is starting with operating system process ID 75914
2024-09-11 13:45:28.626 [ONMD(3629)]CRS-1614: No I/O has completed after
75% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 49730 milliseconds.
2024-09-11 13:45:33.339 [ORAAGENT(76328)]CRS-8500: Oracle Clusterware
ORAAGENT process is starting with operating system process ID 76328
2024-09-11 13:45:58.629 [ONMD(3629)]CRS-1613: No I/O has completed after
90% of the maximum interval. If this persists, voting file
/dev/mapper/grid2 will be considered not functional in 19730 milliseconds.
2024-09-11 13:46:18.630 [ONMD(3629)]CRS-1604: CSSD voting file is offline:
/dev/mapper/grid2; details at (:CSSNM00058:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc.
2024-09-11 13:46:18.631 [ONMD(3629)]CRS-1606: The number of voting files
available, 0, is less than the minimum number of voting files required, 1,
resulting in CSSD termination to ensure data integrity; details at
(:CSSNM00018:) in /gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-11 13:46:18.638 [ONMD(3629)]CRS-1699: The CSS daemon is
terminating due to a fatal error from thread:
clssnmvDiskPingMonitorThread; Details at (:CSSSC00012:) in
/gridbase/diag/crs/jfs13/crs/trace/onmd.trc
2024-09-11 13:46:18.651 [OCSSD(3631)]CRS-1652: Starting clean up of CRSD
resources.
```

O estado do caminho SAN no nó RAC que perdeu os serviços de armazenamento é assim:

```
oradata7 (3600a0980383041334a3f55676c697347) dm-20 NETAPP,LUN C-Mode
size=128G features='3 queue_if_no_path pg_init_retries 50' hwhandler='1
alua' wp=rw
|-+- policy='service-time 0' prio=0 status=enabled
| `- 34:0:0:18 sdam 66:96  failed faulty running
`-+- policy='service-time 0' prio=0 status=enabled
  `- 33:0:0:18 sdaj 66:48  failed faulty running
```

O host linux detetou a perda dos caminhos muito mais rápido do que 200 segundos, mas de uma perspetiva de banco de dados as conexões do cliente com o host no site com falha ainda serão congeladas por 200 segundos sob as configurações padrão do Oracle RAC. As operações de banco de dados completo só serão retomadas após a conclusão do despejo.

Enquanto isso, o nó Oracle RAC no local oposto registrará a perda do outro nó RAC. De outra forma, continua a funcionar como de costume.

```
2024-09-11 13:46:34.152 [ONMD(3547)]CRS-1612: Network communication with
node jfs13 (2) has been missing for 50% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 14.020 seconds
2024-09-11 13:46:41.154 [ONMD(3547)]CRS-1611: Network communication with
node jfs13 (2) has been missing for 75% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 7.010 seconds
2024-09-11 13:46:46.155 [ONMD(3547)]CRS-1610: Network communication with
node jfs13 (2) has been missing for 90% of the timeout interval.  If this
persists, removal of this node from cluster will occur in 2.010 seconds
2024-09-11 13:46:46.470 [OHASD(1705)]CRS-8011: reboot advisory message
from host: jfs13, component: cssmonit, with time stamp: L-2024-09-11-
13:46:46.404
2024-09-11 13:46:46.471 [OHASD(1705)]CRS-8013: reboot advisory message
text: At this point node has lost voting file majority access and
oracssdmonitor is rebooting the node due to unknown reason as it did not
receive local hearbeats for 28180 ms amount of time
2024-09-11 13:46:48.173 [ONMD(3547)]CRS-1632: Node jfs13 is being removed
from the cluster in cluster incarnation 621516934
```

Falha do mediador

O serviço mediador não controla diretamente as operações de storage. Ele funciona como um caminho de controle alternativo entre clusters. Ele existe principalmente para automatizar o failover sem o risco de um cenário de divisão cerebral.

Em operação normal, cada cluster replica alterações em seu parceiro, e cada cluster pode verificar se o cluster do parceiro está on-line e fornecendo dados. Se o link de replicação falhar, a replicação cessaria.

O motivo pelo qual um mediador é necessário para operações automatizadas seguras é porque, de outra forma, seria impossível que os clusters de storage pudessem determinar se a perda de comunicação bidirecional foi o resultado de uma interrupção da rede ou falha real do storage.

O mediador fornece um caminho alternativo para cada cluster para verificar a integridade de seu parceiro. Os cenários são os seguintes:

- Se um cluster puder entrar em Contato diretamente com seu parceiro, os serviços de replicação estarão operacionais. Nenhuma ação necessária.
- Se um site preferido não puder entrar em Contato diretamente com seu parceiro ou por meio do mediador, ele assumirá que ele está realmente indisponível ou foi isolado e que levou seus caminhos LUN off-line. O site preferido continuará lançando o estado RPO/0 e continuará processando e/S de leitura e gravação.
- Se um site não preferencial não puder entrar em Contato diretamente com seu parceiro, mas puder contatá-lo por meio do mediador, ele tomará seus caminhos off-line e aguardará o retorno da conexão de replicação.
- Se um site não preferencial não puder entrar em Contato com seu parceiro diretamente ou por meio de um mediador operacional, ele assumirá que o parceiro está realmente indisponível ou foi isolado e que

tomou seus caminhos LUN off-line. O site não preferencial lançará o estado RPO 0 e continuará processando e/S de leitura e gravação. Ele assumirá o papel da fonte de replicação e se tornará o novo site preferido.

Se o mediador não estiver totalmente disponível:

- A falha dos serviços de replicação por qualquer motivo resultará no lançamento do estado RPO 0 e no reinício do processamento de e/S de leitura e gravação. O site não preferencial tomará seus caminhos off-line.
- A falha do site preferido resultará em uma falha porque o site não-preferido não será capaz de verificar se o site oposto está realmente off-line e, portanto, não seria seguro para o site não-preferido retomar os serviços.

Restauração do serviço

SnapMirror é auto-cura. O SnapMirror active Sync detetará automaticamente a presença de uma relação de replicação com defeito e o levará de volta ao estado RPO igual a 0. Uma vez que a replicação síncrona for restabelecida, os caminhos ficarão online novamente.

Em muitos casos, os aplicativos em cluster detetarão automaticamente o retorno de caminhos com falha, e esses aplicativos também voltarão online. Em outros casos, pode ser necessária uma análise SAN no nível do host ou os aplicativos podem precisar ser colocados online manualmente.

Depende do aplicativo e de como ele é configurado e, em geral, essas tarefas podem ser facilmente automatizadas. O SnapMirror active Sync em si é auto-consertado e não deve exigir a intervenção do usuário para retomar as operações de storage RPO 0 depois que a energia e a conectividade forem restauradas.

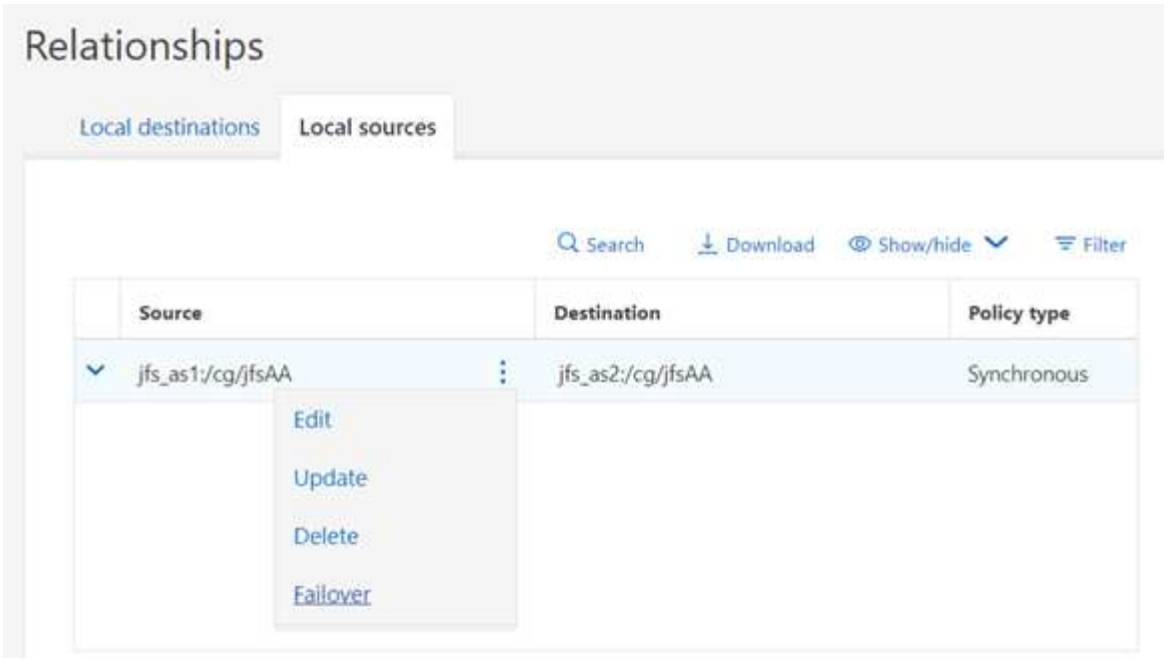
Failover manual

O termo "failover" não se refere à direção da replicação com a sincronização ativa do SnapMirror porque é uma tecnologia de replicação bidirecional. Em vez disso, "failover" refere-se a qual sistema de armazenamento será o local preferido em caso de falha.

Por exemplo, você pode querer executar um failover para alterar o local preferido antes de encerrar um local para manutenção ou antes de executar um teste de DR.

Alterar o local preferido requer uma operação simples. A e/S pausa por um segundo ou dois como autoridade sobre os switches de comportamento de replicação entre clusters, mas a e/S não é afetada.

Exemplo de GUI:



Exemplo de alterá-lo de volta através da CLI:

```
Cluster2::> snapmirror failover start -destination-path jfs_as2:/cg/jfsAA
[Job 9575] Job is queued: SnapMirror failover for destination
"jfs_as2:/cg/jfsAA                        ".

Cluster2::> snapmirror failover show

Source    Destination                                            Error
Path      Path        Type     Status    start-time end-time    Reason
--------- ----------  -------- --------- ---------- ---------- ----------
jfs_as1:/cg/jfsAA
          jfs_as2:/cg/jfsAA
                      planned  completed 9/11/2024  9/11/2024
                                         09:29:22   09:29:32

The new destination path can be verified as follows:


Cluster1::> snapmirror show -destination-path jfs_as1:/cg/jfsAA

                           Source Path: jfs_as2:/cg/jfsAA
                        Destination Path: jfs_as1:/cg/jfsAA
                       Relationship Type: XDP
                 Relationship Group Type: consistencygroup
                  SnapMirror Policy Type: automated-failover-duplex
                       SnapMirror Policy: AutomatedFailOverDuplex
                             Tries Limit: -
                           Mirror State: Snapmirrored
                    Relationship Status: InSync
```

Migração de banco de dados Oracle

Visão geral

Aproveitar os recursos de uma nova plataforma de storage tem um requisito inevitável; os dados devem ser colocados no novo sistema de storage. O ONTAP simplifica o processo de migração, incluindo migrações e atualizações do ONTAP para o ONTAP, importações de LUN estrangeiras e procedimentos para o uso do sistema operacional host ou do software de banco de dados Oracle diretamente.

Esta documentação substitui o relatório técnico publicado anteriormente *TR-4534: Migração de bancos de dados Oracle para sistemas de armazenamento NetApp*

No caso de um novo projeto de banco de dados, isso não é uma preocupação porque os ambientes de banco de dados e aplicativos são construídos no lugar. A migração, no entanto, apresenta desafios especiais em relação à interrupção dos negócios, ao tempo necessário para a conclusão da migração, aos conjuntos de

habilidades necessários e à minimização dos riscos.

Scripts

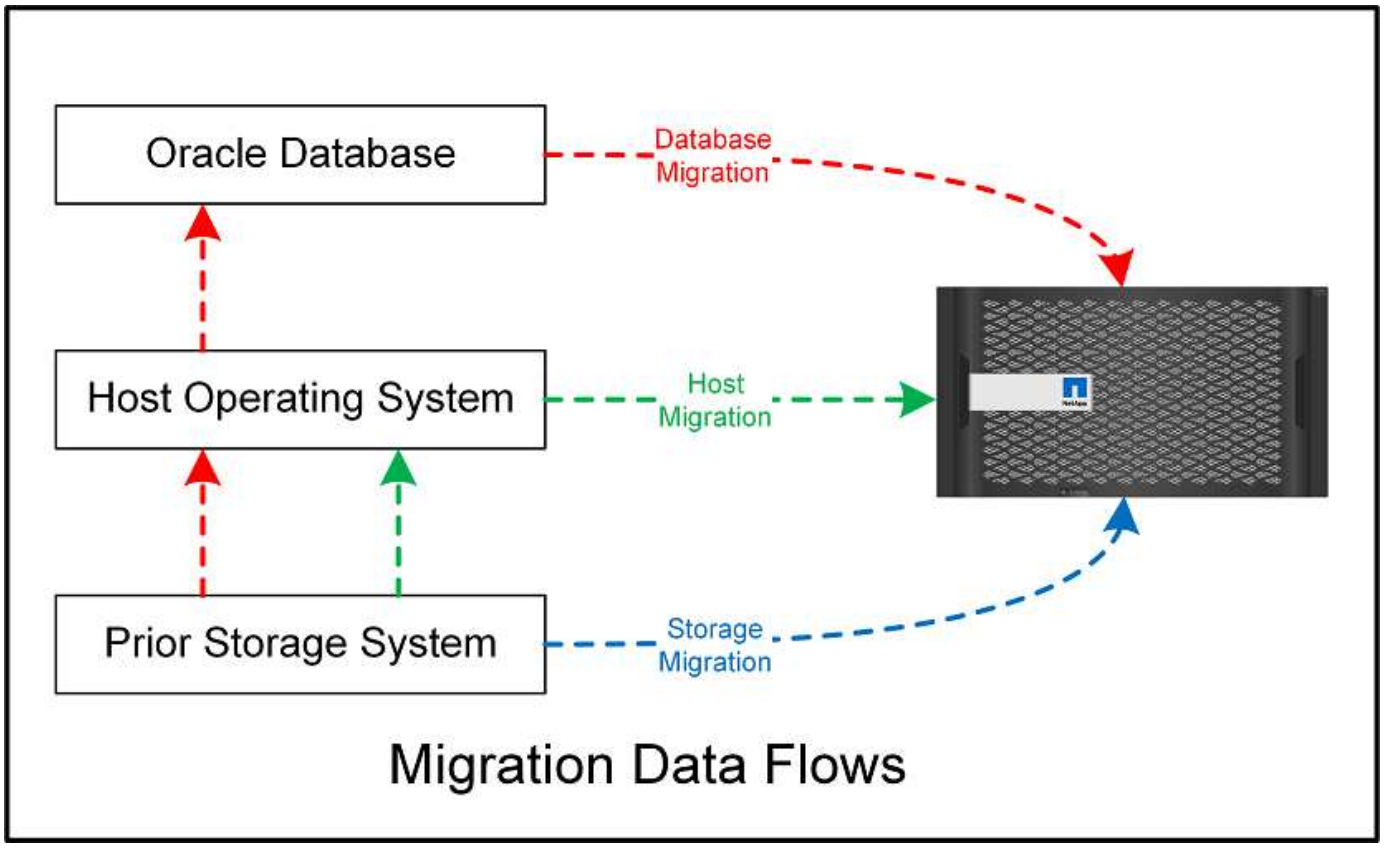
Exemplos de scripts são fornecidos nesta documentação. Esses scripts fornecem métodos de exemplo para automatizar vários aspetos da migração para reduzir a chance de erros do usuário. Os scripts podem reduzir as demandas gerais da equipe DE TI responsável por uma migração e acelerar o processo geral. Todos esses scripts são extraídos de projetos de migração reais executados pelos Serviços profissionais da NetApp e parceiros da NetApp. Exemplos de seu uso são mostrados ao longo desta documentação.

Planejamento de migração

A migração de dados Oracle pode ocorrer em um de três níveis: O banco de dados, o host ou o storage array.

As diferenças estão em qual componente da solução geral é responsável pela movimentação de dados: O banco de dados, o sistema operacional host ou o sistema de storage.

A figura abaixo mostra um exemplo dos níveis de migração e do fluxo de dados. No caso da migração no nível do banco de dados, os dados são movidos do sistema de storage original pelas camadas do host e do banco de dados para o novo ambiente. A migração em nível de host é semelhante, mas os dados não passam pela camada de aplicativo e são gravados no novo local usando processos de host. Finalmente, com a migração no nível do storage, um array como um sistema NetApp FAS é responsável pela movimentação de dados.



Uma migração no nível do banco de dados geralmente se refere ao uso do Oracle log enviado por meio de um banco de dados de reserva para concluir uma migração na camada Oracle. As migrações em nível de host são executadas usando a capacidade nativa da configuração do sistema operacional do host. Essa configuração inclui operações de cópia de arquivos usando comandos como cp, tar e Oracle Recovery Manager (RMAN) ou usando um gerenciador de volume lógico (LVM) para realocar os bytes subjacentes de

um sistema de arquivos. O Oracle Automatic Storage Management (ASM) é categorizado como um recurso no nível do host, porque é executado abaixo do nível do aplicativo de banco de dados. O ASM ocupa o lugar do gerenciador de volume lógico usual em um host. Finalmente, os dados podem ser migrados no nível de storage array, o que significa que estão abaixo do nível do sistema operacional.

Considerações de Planejamento

A melhor opção para migração depende de uma combinação de fatores, incluindo a escala do ambiente a ser migrado, a necessidade de evitar tempo de inatividade e o esforço geral necessário para realizar a migração. Bancos de dados grandes obviamente exigem mais tempo e esforço para a migração, mas a complexidade dessa migração é mínima. Pequenos bancos de dados podem ser migrados rapidamente, mas, se houver milhares para serem migrados, a escala do esforço pode criar complicações. Finalmente, quanto maior o banco de dados, maior a probabilidade de ser crítico para os negócios, o que dá origem a uma necessidade de minimizar o tempo de inatividade, preservando um caminho de back-out.

Algumas das considerações para o Planejamento de uma estratégia de migração são discutidas aqui.

Tamanho dos dados

Os tamanhos dos bancos de dados a serem migrados obviamente afetam o Planejamento da migração, embora o tamanho não afete necessariamente o tempo de transição. Quando uma grande quantidade de dados precisa ser migrada, a principal consideração é a largura de banda. As operações de cópia são geralmente realizadas com e/S sequenciais eficientes Como uma estimativa conservadora, suponha 50% de utilização da largura de banda de rede disponível para operações de cópia. Por exemplo, uma porta FC de 8GB GB pode transferir cerca de 800MBps GB teoricamente. Assumindo 50% de utilização, um banco de dados pode ser copiado a uma taxa de cerca de 400Mbps. Portanto, um banco de dados 10TB pode ser copiado em cerca de sete horas a essa taxa.

A migração em distâncias maiores geralmente requer uma abordagem mais criativa, como o processo de envio de logs explicado em "Movimentação online do arquivo de dados". As redes IP de longa distância raramente têm largura de banda em qualquer lugar perto das velocidades LAN ou SAN. Em um caso, o NetApp ajudou com a migração de longa distância de um banco de dados 220TB com taxas de geração de arquivos log muito altas. A abordagem escolhida para transferência de dados foi o envio diário de fitas, pois esse método oferecia a máxima largura de banda possível.

Contagem da base de dados

Em muitos casos, o problema de mover uma grande quantidade de dados não é o tamanho dos dados, mas sim a complexidade da configuração que suporta o banco de dados. Simplesmente saber que 50TB de bancos de dados devem ser migrados não é informação suficiente. Pode ser um único banco de dados de missão crítica 50TB, uma coleção de 4, 000 bancos de dados legados ou uma combinação de dados de produção e não produção. Em alguns casos, grande parte dos dados consiste em clones de um banco de dados de origem. Esses clones não precisam ser migrados porque podem ser facilmente recriados, especialmente quando a nova arquitetura foi projetada para utilizar volumes NetApp FlexClone.

Para Planejar a migração, você deve entender quantos bancos de dados estão no escopo e como eles devem ser priorizados. À medida que o número de bancos de dados aumenta, a opção de migração preferida tende a ser menor e menor na pilha. Por exemplo, copiar um único banco de dados pode ser facilmente executado com RMAN e uma pequena interrupção. Essa é a replicação no nível do host.

Se houver bancos de dados 50, pode ser mais fácil evitar a configuração de uma nova estrutura de sistema de arquivos para receber uma cópia RMAN e, em vez disso, mover os dados no lugar. Esse processo pode ser feito aproveitando a migração LVM baseada em host para realocar dados de LUNs antigos para novos LUNs. Isso move a responsabilidade da equipe do administrador do banco de dados (DBA) para a equipe do sistema operacional e, como resultado, os dados são migrados de forma transparente em relação ao banco de dados.

A configuração do sistema de arquivos não foi alterada.

Por fim, se for necessário migrar bancos de dados 500 em servidores 200, opções baseadas em armazenamento, como o recurso de importação de LUN estrangeiro (FLI) do ONTAP, podem ser usadas para realizar uma migração direta dos LUNs.

Requisitos de rearquitetura

Normalmente, um layout de arquivo de banco de dados deve ser alterado para aproveitar os recursos do novo storage array; no entanto, isso nem sempre é o caso. Por exemplo, os recursos dos all-flash arrays EF-Series são direcionados principalmente à performance de SAN e à confiabilidade de SAN. Na maioria dos casos, os bancos de dados podem ser migrados para um array da EF-Series sem considerações especiais para o layout de dados. Os únicos requisitos são alto IOPS, baixa latência e confiabilidade robusta. Embora existam práticas recomendadas relacionadas a fatores como configuração RAID ou Dynamic Disk Pools, os projetos EF-Series raramente exigem alterações significativas na arquitetura de storage geral para utilizar esses recursos.

Em contraste, a migração para o ONTAP geralmente requer mais consideração do layout do banco de dados para garantir que a configuração final forneça o valor máximo. Por si só, o ONTAP oferece muitos recursos para um ambiente de banco de dados, mesmo sem nenhum esforço de arquitetura específico. O mais importante é que ele oferece a capacidade de migrar sem interrupções para um novo hardware quando o hardware atual atinge seu fim de vida útil. De um modo geral, uma migração para o ONTAP é a última migração que você precisaria executar. O hardware subsequente é atualizado e os dados são migrados para novas Mídias sem interrupções.

Com algum Planejamento, ainda mais benefícios estão disponíveis. As considerações mais importantes envolvem o uso de instantâneos. Os snapshots são a base para realizar backups, restaurações e operações de clonagem quase instantâneos. Como exemplo do poder dos snapshots, o maior uso conhecido é com um único banco de dados de 996TB GB executado em cerca de 250 LUNs em controladores de 6. Esse banco de dados pode ser feito em 2 minutos, restaurado em 2 minutos e clonado em 15 minutos. Os benefícios adicionais incluem a capacidade de mover dados pelo cluster em resposta a alterações no workload e a aplicação de controles de qualidade do serviço (QoS) para fornecer performance boa e consistente em um ambiente de vários bancos de dados.

Tecnologias como controles de QoS, realocação de dados, snapshots e clonagem funcionam em praticamente qualquer configuração. No entanto, alguns pensamentos geralmente são necessários para maximizar os benefícios. Em alguns casos, os layouts de armazenamento de banco de dados podem exigir alterações de design para maximizar o investimento no novo storage array. Essas alterações no design podem afetar a estratégia de migração porque as migrações baseadas em host ou baseadas em storage replicam o layout de dados original. Etapas adicionais podem ser necessárias para concluir a migração e fornecer um layout de dados otimizado para o ONTAP. Os procedimentos mostrados em "Visão geral dos procedimentos de migração Oracle" e mais tarde demonstram alguns dos métodos para não apenas migrar um banco de dados, mas para migrá-lo para o layout final ideal com o mínimo de esforço.

Tempo de redução

A interrupção máxima permitida de serviço durante a transição deve ser determinada. É um erro comum supor que todo o processo de migração causa interrupções. Muitas tarefas podem ser concluídas antes do início de qualquer interrupção de serviço e muitas opções permitem a conclusão da migração sem interrupção ou interrupção. Mesmo quando a interrupção é inevitável, você ainda deve definir a interrupção máxima de serviço permitida, pois a duração do tempo de transição varia de procedimento para procedimento.

Por exemplo, copiar um banco de dados 10TB normalmente requer aproximadamente sete horas para ser concluído. Se as necessidades empresariais permitirem uma interrupção de sete horas, a cópia de arquivos é uma opção fácil e segura para migração. Se cinco horas forem inaceitáveis, um processo simples de envio de

logs (consulte "Oracle log de envio") pode ser configurado com o mínimo de esforço para reduzir o tempo de transferência para aproximadamente 15 minutos. Durante esse tempo, um administrador de banco de dados pode concluir o processo. Se 15 minutos forem inaceitáveis, o processo de transição final pode ser automatizado por meio de script para reduzir o tempo de transição para apenas alguns minutos. Você sempre pode acelerar uma migração, mas isso acontece com o custo de tempo e esforço. As metas de tempo de transição devem ser baseadas no que é aceitável para o negócio.

Caminho de back-out

Nenhuma migração é completamente livre de riscos. Mesmo que a tecnologia funcione perfeitamente, há sempre a possibilidade de erro do usuário. O risco associado a um caminho de migração escolhido deve ser considerado juntamente com as consequências de uma migração falhada. Por exemplo, o recurso transparente de migração de armazenamento on-line do Oracle ASM é um de seus principais recursos, e esse método é um dos mais confiáveis conhecidos. No entanto, os dados estão sendo copiados irreversivelmente com este método. No caso altamente improvável de que um problema ocorra com ASM, não há um caminho de back-out fácil. A única opção é restaurar o ambiente original ou usar o ASM para reverter a migração de volta para os LUNs originais. O risco pode ser minimizado, mas não eliminado, executando um backup do tipo snapshot no sistema de storage original, supondo que o sistema seja capaz de executar tal operação.

Ensaio

Alguns procedimentos de migração devem ser totalmente verificados antes da execução. A necessidade de migração e ensaio do processo de transição é uma solicitação comum com bancos de dados de missão crítica para os quais a migração deve ser bem-sucedida e o tempo de inatividade deve ser minimizado. Além disso, os testes de aceitação do usuário são frequentemente incluídos como parte do trabalho de pós-migração, e o sistema geral pode ser devolvido à produção somente após a conclusão desses testes.

Se houver necessidade de ensaio, vários recursos do ONTAP podem tornar o processo muito mais fácil. Em particular, os snapshots podem redefinir um ambiente de teste e criar rapidamente várias cópias com uso eficiente de espaço de um ambiente de banco de dados.

Procedimentos

Visão geral

Muitos procedimentos estão disponíveis para o banco de dados de migração Oracle. O certo depende das necessidades do seu negócio.

Em muitos casos, os administradores de sistema e DBAs têm seus próprios métodos preferidos de relocalização de dados de volume físico, espelhamento e desirritações, ou de aproveitamento do Oracle RMAN para copiar dados.

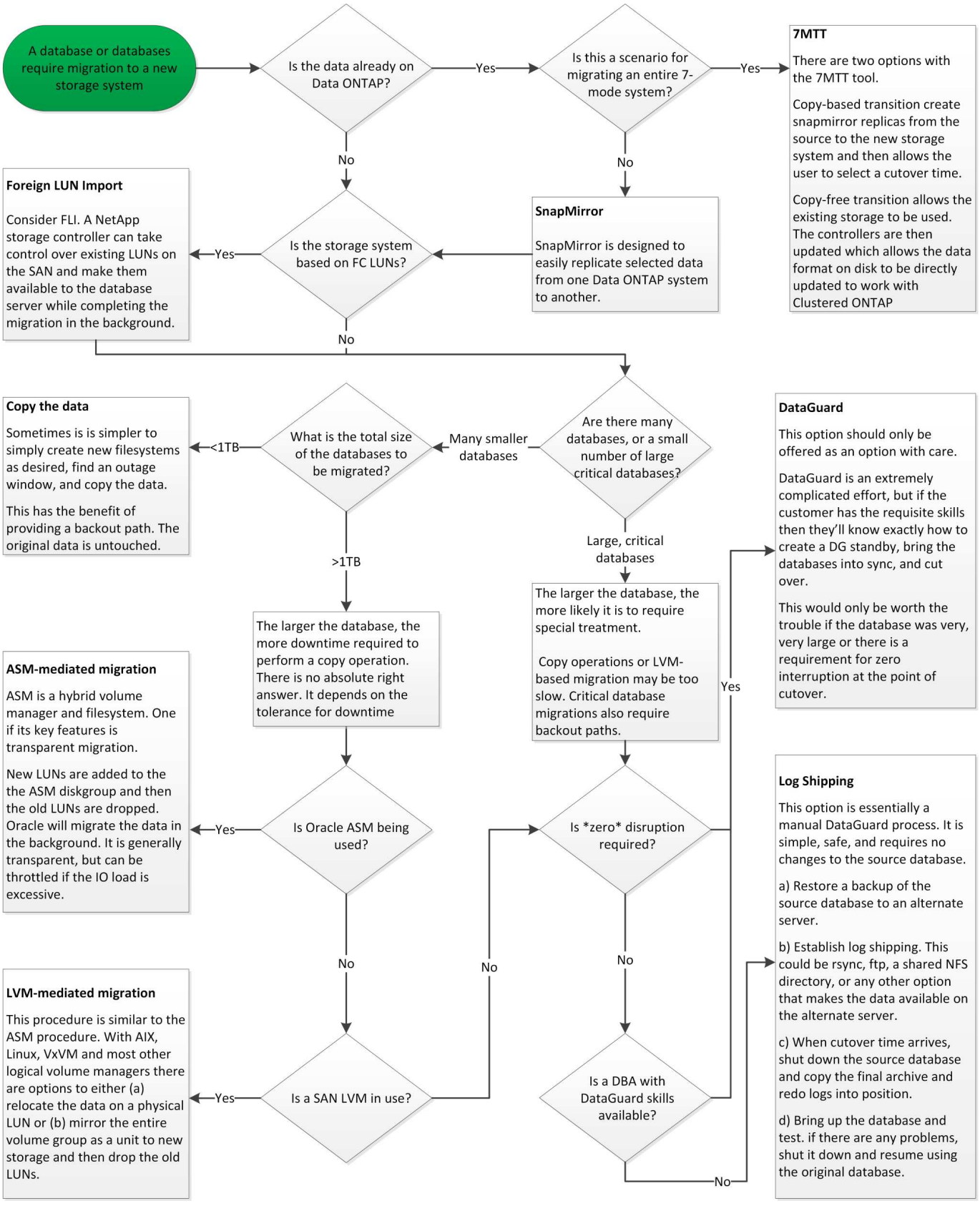
Esses procedimentos são fornecidos principalmente como orientação para a equipe DE TI menos familiarizada com algumas das opções disponíveis. Além disso, os procedimentos ilustram as tarefas, os requisitos de tempo e as demandas do conjunto de habilidades para cada abordagem de migração. Isso permite que outras partes, como NetApp e serviços profissionais de parceiros ou gerenciamento DE TI, apreciem mais plenamente os requisitos de cada procedimento.

Não existe uma prática recomendada única para a criação de uma estratégia de migração. A criação de um plano requer primeiro compreender as opções de disponibilidade e, em seguida, selecionar o método que melhor se adapta às necessidades do negócio. A figura abaixo ilustra as considerações básicas e conclusões típicas feitas pelos clientes, mas não é universalmente aplicável a todas as situações.

Por exemplo, uma etapa levanta a questão do tamanho total do banco de dados. A próxima etapa depende se

o banco de dados é mais ou menos do que 1TB. As etapas recomendadas são exatamente isso: Recomendações baseadas em práticas típicas do cliente. A maioria dos clientes não usaria o DataGuard para copiar um pequeno banco de dados, mas alguns podem. A maioria dos clientes não tentaria copiar um banco de dados 50TB devido ao tempo necessário, mas alguns podem ter uma janela de manutenção suficientemente grande para permitir tal operação.

O fluxograma abaixo mostra os tipos de considerações sobre qual caminho de migração é melhor. Você pode clicar com o botão direito na imagem e abri-la em uma nova guia para melhorar a legibilidade.



.

Movimentação online do arquivo de dados

Oracle 12cR1 e superior incluem a capacidade de mover um arquivo de dados enquanto o banco de dados permanece online. Ele também funciona entre diferentes tipos de sistema de arquivos. Por exemplo, um

arquivo de dados pode ser relocado de um sistema de arquivos xfs para ASM. Esse método geralmente não é usado em escala devido ao número de operações individuais de movimentação de arquivos de dados que seriam necessárias, mas é uma opção que vale a pena considerar com bancos de dados menores com menos datafiles.

Além disso, simplesmente mover um arquivo de dados é uma boa opção para migrar partes de bancos de dados existentes. Por exemplo, datafiles menos ativos podem ser relocados para um storage mais econômico, como um volume FabricPool que pode armazenar blocos ociosos no armazenamento de objetos.

Migração em nível de banco de dados

Migração no nível do banco de dados significa permitir que o banco de dados relocate os dados. Especificamente, isso significa o envio de logs. Tecnologias como RMAN e ASM são produtos Oracle, mas, para fins de migração, elas operam no nível do host, onde copiam arquivos e gerenciam volumes.

Registrar envio

A base para a migração no nível do banco de dados é o log de arquivo Oracle, que contém um log de alterações no banco de dados. Na maioria das vezes, um log de arquivamento faz parte de uma estratégia de backup e recuperação. O processo de recuperação começa com a restauração de um banco de dados e, em seguida, a reprodução de um ou mais logs de arquivo para trazer o banco de dados para o estado desejado. Essa mesma tecnologia básica pode ser usada para realizar uma migração com pouca ou nenhuma interrupção das operações. Mais importante ainda, essa tecnologia permite a migração ao mesmo tempo em que deixa o banco de dados original intocado, preservando um caminho de back-out.

O processo de migração começa com a restauração de um backup de banco de dados para um servidor secundário. Você pode fazer isso de várias maneiras, mas a maioria dos clientes usa seu aplicativo de backup normal para restaurar os arquivos de dados. Depois que os arquivos de dados são restaurados, os usuários estabelecem um método para o envio de log. O objetivo é criar um feed constante de logs de arquivo gerados pelo banco de dados principal e reproduzi-los no banco de dados restaurado para mantê-los ambos perto do mesmo estado. Quando o tempo de transição chega, o banco de dados de origem é completamente desligado e os Registros finais do arquivo e, em alguns casos, os logs de refazer são copiados e reproduzidos. É fundamental que os logs de refazer também sejam considerados porque eles podem conter algumas das transações finais confirmadas.

Depois que esses logs foram transferidos e reproduzidos, ambos os bancos de dados são consistentes uns com os outros. Neste ponto, a maioria dos clientes realiza alguns testes básicos. Se algum erro for feito durante o processo de migração, a repetição do log deve relatar erros e falhar. Ainda é aconselhável realizar alguns testes rápidos com base em consultas conhecidas ou atividades orientadas por aplicativos para verificar se a configuração é ideal. Também é uma prática comum criar uma tabela de teste final antes de encerrar o banco de dados original para verificar se ele está presente no banco de dados migrado. Esta etapa garante que não foram feitos erros durante a sincronização final do log.

Uma simples migração de envio de logs pode ser configurada fora da banda em relação ao banco de dados original, o que o torna particularmente útil para bancos de dados de missão crítica. Nenhuma alteração de configuração é necessária para o banco de dados de origem, e a restauração e configuração inicial do ambiente de migração não têm efeito sobre as operações de produção. Depois que o envio de log é configurado, ele coloca algumas demandas de e/S nos servidores de produção. No entanto, o envio de logs consiste em leituras sequenciais simples dos logs do arquivo, o que é improvável que tenha qualquer efeito no desempenho do banco de dados de produção.

O envio de logs provou ser particularmente útil para projetos de migração de longa distância e alta taxa de mudança. Em um caso, um único banco de dados 220TB foi migrado para um novo local a aproximadamente 500 km de distância. A taxa de mudança foi extremamente alta e as restrições de segurança impediram o uso de uma conexão de rede. O envio de log foi realizado usando fita adesiva e correio. Uma cópia do banco de

dados de origem foi inicialmente restaurada usando os procedimentos descritos abaixo. Os logs foram então enviados semanalmente pelo correio até o momento da transição quando o conjunto final de fitas foi entregue e os logs foram aplicados ao banco de dados de réplica.

Oracle DataGuard

Em alguns casos, um ambiente DataGuard completo é garantido. É incorreto usar o termo DataGuard para se referir a qualquer envio de log ou configuração de banco de dados em espera. O Oracle DataGuard é uma estrutura abrangente para gerenciar a replicação de banco de dados, mas não é uma tecnologia de replicação. O principal benefício de um ambiente DataGuard completo em um esforço de migração é o switchover transparente de um banco de dados para outro. O DataGuard também permite um switchover transparente de volta para o banco de dados original se um problema for descoberto, como um problema de desempenho ou conetividade de rede com o novo ambiente. Um ambiente DataGuard totalmente configurado requer a configuração não apenas da camada de banco de dados, mas também de aplicativos para que os aplicativos sejam capazes de detetar uma alteração na localização do banco de dados primário. Em geral, não é necessário usar o DataGuard para concluir uma migração, mas alguns clientes têm uma vasta experiência do DataGuard internamente e já dependem dele para o trabalho de migração.

Rearquitetura

Como discutido anteriormente, aproveitar os recursos avançados de storage arrays às vezes requer a alteração do layout do banco de dados. Além disso, uma alteração no protocolo de armazenamento, como a mudança de ASM para um sistema de arquivos NFS, necessariamente altera o layout do sistema de arquivos.

Uma das principais vantagens dos métodos de envio de log, incluindo o DataGuard, é que o destino de replicação não precisa corresponder à origem. Não há problemas com o uso de uma abordagem de envio de logs para migrar do ASM para um sistema de arquivos regular ou vice-versa. O layout preciso dos arquivos de dados pode ser alterado no destino para otimizar o uso da tecnologia PDB (Pluggable Database) ou para definir controles de QoS seletivamente em determinados arquivos. Em outras palavras, um processo de migração baseado no envio de logs permite otimizar o layout de armazenamento de banco de dados de forma fácil e segura.

Recursos do servidor

Uma limitação à migração no nível do banco de dados é a necessidade de um segundo servidor. Há duas maneiras de usar este segundo servidor:

1. Você pode usar o segundo servidor como uma nova casa permanente para o banco de dados.
2. Você pode usar o segundo servidor como um servidor de teste temporário. Depois que a migração de dados para o novo storage array for concluída e testada, os sistemas de arquivos LUN ou NFS são desconetados do servidor de teste e reconetados ao servidor original.

A primeira opção é a mais fácil, mas usá-la pode não ser viável em ambientes muito grandes que exigem servidores muito poderosos. A segunda opção requer trabalho extra para realocar os sistemas de arquivos de volta para o local original. Essa pode ser uma operação simples na qual o NFS é usado como protocolo de armazenamento, pois os sistemas de arquivos podem ser desmontados do servidor de teste e remontados no servidor original.

Os sistemas de arquivos baseados em blocos exigem trabalho extra para atualizar o zoneamento FC ou iniciadores iSCSI. Com a maioria dos gerenciadores lógicos de volume (incluindo ASM), os LUNs são detetados e colocados on-line automaticamente depois que são disponibilizados no servidor original. No entanto, algumas implementações de sistema de arquivos e LVM podem exigir mais trabalho para exportar e importar os dados. O procedimento preciso pode variar, mas geralmente é fácil estabelecer um procedimento simples e repetível para concluir a migração e realojar os dados no servidor original.

Embora seja possível configurar o envio de log e replicar um banco de dados em um único ambiente de servidor, a nova instância deve ter um SID de processo diferente para reproduzir os logs. É possível abrir temporariamente o banco de dados em um conjunto diferente de IDs de processo com um SID diferente e alterá-lo mais tarde. No entanto, isso pode levar a muitas atividades de gerenciamento complicadas, e coloca o ambiente de banco de dados em risco de erro do usuário.

Migração em nível de host

Migrar dados no nível do host significa usar o sistema operacional host e os utilitários associados para concluir a migração. Esse processo inclui qualquer utilitário que copia dados, incluindo Oracle RMAN e Oracle ASM.

Cópia de dados

O valor de uma operação de cópia simples não deve ser subestimado. Infraestruturas de rede modernas podem mover dados a taxas medidas em gigabytes por segundo, e as operações de cópia de arquivos são baseadas em e/S de leitura e gravação sequencial eficiente Mais interrupções são inevitáveis com uma operação de cópia de host quando comparada ao envio de logs, mas uma migração é mais do que apenas a movimentação de dados. Geralmente inclui alterações na rede, no tempo de reinicialização do banco de dados e nos testes de pós-migração.

O tempo real necessário para copiar dados pode não ser significativo. Além disso, uma operação de cópia preserva um caminho de back-out garantido porque os dados originais permanecem intocados. Se algum problema for encontrado durante o processo de migração, os sistemas de arquivos originais com os dados originais podem ser reativados.

Replatforming

Replatforming refere-se a uma alteração no tipo de CPU. Quando um banco de dados é migrado de uma plataforma tradicional Solaris, AIX ou HP-UX para o Linux x86, os dados devem ser reformatados devido a alterações na arquitetura da CPU. SPARC, IA64 e CPUs DE ENERGIA são conhecidos como processadores big endian, enquanto as arquiteturas x86 e x86_64 são conhecidas como little endian. Como resultado, alguns dados dentro dos arquivos de dados Oracle são ordenados de forma diferente, dependendo do processador em uso.

Tradicionalmente, os clientes usam o DataPump para replicar dados entre plataformas. DataPump é um utilitário que cria um tipo especial de exportação de dados lógicos que pode ser mais rapidamente importado no banco de dados de destino. Como ele cria uma cópia lógica dos dados, o DataPump deixa as dependências da endianidade do processador para trás. O DataPump ainda é usado por alguns clientes para replatforming, mas uma opção mais rápida se tornou disponível com o Oracle 11g: Tablespaces transportáveis entre plataformas. Este avanço permite que um espaço de tablespace seja convertido para um formato de endian diferente no lugar. Esta é uma transformação física que oferece melhor desempenho do que uma exportação DataPump, que deve converter bytes físicos em dados lógicos e depois converter de volta para bytes físicos.

Uma discussão completa sobre DataPump e tablespaces transportáveis está além da documentação do Scope NetApp, mas o NetApp tem algumas recomendações com base em nossa experiência ajudando os clientes durante a migração para um novo log de storage array com uma nova arquitetura de CPU:

- Se o DataPump estiver sendo usado, o tempo necessário para concluir a migração deve ser medido em um ambiente de teste. Às vezes, os clientes ficam surpresos no momento necessário para concluir a migração. Este tempo de inatividade adicional inesperado pode causar interrupções.
- Muitos clientes acreditam erroneamente que as tabelas transportáveis entre plataformas não exigem conversão de dados. Quando uma CPU com um endian diferente é usada, uma operação RMAN

`convert` deve ser executada nos arquivos de dados de antemão. Esta não é uma operação instantânea. Em alguns casos, o processo de conversão pode ser acelerado por ter vários threads operando em diferentes arquivos de dados, mas o processo de conversão não pode ser evitado.

Migração lógica orientada pelo Gerenciador de volumes

Os LVMs funcionam tomando um grupo de uma ou mais LUNs e dividindo-os em pequenas unidades geralmente chamadas de extensões. O conjunto de extensões é então usado como uma fonte para criar volumes lógicos que são essencialmente virtualizados. Essa camada de virtualização agrega valor de várias maneiras:

- Volumes lógicos podem usar extensões desenhadas a partir de vários LUNs. Quando um sistema de arquivos é criado em um volume lógico, ele pode usar todas as funcionalidades de performance de todos os LUNs. Ele também promove o carregamento uniforme de todos os LUNs no grupo de volumes, fornecendo performance mais previsível.
- Os volumes lógicos podem ser redimensionados adicionando e, em alguns casos, removendo extensões. Redimensionar um sistema de arquivos em um volume lógico geralmente não causa interrupções.
- Os volumes lógicos podem ser migrados sem interrupções ao mover as extensões subjacentes.

A migração usando um LVM funciona de duas maneiras: Mover uma extensão ou espelhar/desirritar uma extensão. A migração LVM usa e/S sequenciais de grandes blocos eficientes e raramente cria preocupações de desempenho. Se isso se tornar um problema, geralmente há opções para limitar a taxa de e/S. Isso aumenta o tempo necessário para concluir a migração e, ao mesmo tempo, reduz a sobrecarga de e/S nos sistemas de host e storage.

Espelho e demirror

Alguns gerenciadores de volume, como o AIX LVM, permitem que o usuário especifique o número de cópias para cada extensão e controle quais dispositivos hospedam cada cópia. A migração é feita pegando um volume lógico existente, espelhando as extensões subjacentes aos novos volumes, aguardando a sincronização das cópias e, em seguida, deixando cair a cópia antiga. Se um caminho de saída for desejado, um instantâneo dos dados originais pode ser criado antes do ponto em que a cópia espelhada é descartada. Como alternativa, o servidor pode ser encerrado brevemente para mascarar LUNs originais antes de excluir forçosamente as cópias espelhadas contidas. Isso preserva uma cópia recuperável dos dados em seu local original.

Migração de extensão

Quase todos os gerenciadores de volume permitem que extensões sejam migradas e, às vezes, existem várias opções. Por exemplo, alguns gerenciadores de volume permitem que um administrador relocate as extensões individuais para um volume lógico específico do armazenamento antigo para o novo. Os gerenciadores de volume, como o Linux LVM2, oferecem o `pvmove` comando, que relocaliza todas as extensões no dispositivo LUN especificado para um novo LUN. Depois que o LUN antigo é evacuado, ele pode ser removido.

O principal risco para as operações é a remoção de LUNs antigos e não utilizados da configuração. É preciso ter muito cuidado ao alterar o zoneamento FC e remover dispositivos LUN obsoletos.

Gerenciamento automático de armazenamento Oracle

O Oracle ASM é um gerenciador de volumes lógicos e um sistema de arquivos combinados. Em um alto nível, o Oracle ASM toma uma coleção de LUNs, os divide em pequenas unidades de alocação e os apresenta

como um único volume conhecido como um grupo de discos ASM. O ASM também inclui a capacidade de espelhar o grupo de discos definindo o nível de redundância. Um volume pode ser sem espelhamento (redundância externa), espelhado (redundância normal) ou espelhado em três vias (redundância alta). É necessário ter cuidado ao configurar o nível de redundância porque não pode ser alterado após a criação.

O ASM também fornece a funcionalidade do sistema de arquivos. Embora o sistema de arquivos não seja visível diretamente do host, o banco de dados Oracle pode criar, mover e excluir arquivos e diretórios em um grupo de discos ASM. Além disso, a estrutura pode ser navegada usando o utilitário asmcmd.

Assim como em outras implementações de LVM, o Oracle ASM otimiza a performance de e/S por meio da distribuição e balanceamento de carga da e/S de cada arquivo em todas as LUNs disponíveis. Em segundo lugar, as extensões subjacentes podem ser relocadas para permitir o redimensionamento do grupo de discos ASM, bem como a migração. O Oracle ASM automatiza o processo por meio da operação de rebalanceamento. Novos LUNs são adicionados a um grupo de discos ASM e os LUNs antigos são descartados, o que aciona a realocação de extensão e a subsequente queda do LUN evacuado do grupo de discos. Esse processo é um dos métodos de migração mais comprovados, e a confiabilidade do ASM na entrega de migração transparente é possivelmente sua caraterística mais importante.

Como o nível de espelhamento do Oracle ASM é fixo, ele não pode ser usado com o método de migração mirror e Demirror.

Migração no nível de storage

Migração no nível de storage significa realizar a migração abaixo do nível de aplicativo e do sistema operacional. No passado, isso às vezes significava usar dispositivos especializados que copiavam LUNs no nível da rede, mas esses recursos agora são encontrados nativamente no ONTAP.

SnapMirror

A migração de bancos de dados entre sistemas NetApp é quase universalmente realizada com o software de replicação de dados NetApp SnapMirror. O processo envolve a configuração de uma relação de espelho para os volumes a serem migrados, permitindo que eles sincronizem e, em seguida, aguardando a janela de transição. Quando ele chega, o banco de dados de origem é desligado, uma atualização final do espelho é executada e o espelho é quebrado. Os volumes de réplica ficam prontos para uso, seja pela montagem de um diretório de sistema de arquivos NFS contido ou descobrindo os LUNs contidos e iniciando o banco de dados.

A realocação de volumes em um único cluster do ONTAP não é considerada migração, mas sim uma operação de rotina `volume move`. O SnapMirror é usado como o mecanismo de replicação de dados no cluster. Este processo é totalmente automatizado. Não há etapas adicionais de migração a serem executadas quando os atributos do volume, como mapeamento LUN ou permissões de exportação NFS, são movidos com o próprio volume. A realocação não causa interrupções nas operações de host. Em alguns casos, o acesso à rede deve ser atualizado para garantir que os dados recém-realocados sejam acessados da maneira mais eficiente possível, mas essas tarefas também não causam interrupções.

Importação de LUN estrangeiro (FLI)

O FLI é um recurso que permite que um sistema Data ONTAP executando 8,3 ou superior migre um LUN existente de outro storage array. O procedimento é simples: O sistema ONTAP está localizado no storage array existente como se fosse qualquer outro host SAN. O Data ONTAP então assume o controle dos LUNs legados desejados e migra os dados subjacentes. Além disso, o processo de importação usa as configurações de eficiência do novo volume à medida que os dados são migrados, o que significa que os dados podem ser compatados e desduplicados em linha durante o processo de migração.

A primeira implementação do FLI no Data ONTAP 8.3 permitiu apenas migração off-line. Esta foi uma

transferência extremamente rápida, mas ainda significava que os dados LUN estavam indisponíveis até que a migração fosse concluída. A migração online foi introduzida no Data ONTAP 8.3,1. Esse tipo de migração minimiza a interrupção ao permitir que o ONTAP atenda dados LUN durante o processo de transferência. Há uma breve interrupção, enquanto o host é rezoneado para usar os LUNs por meio do ONTAP. No entanto, assim que essas alterações forem feitas, os dados serão novamente acessíveis e permanecem acessíveis durante todo o processo de migração.

A leitura de e/S é suportada através do ONTAP até que a operação de cópia esteja concluída, enquanto a escrita de e/S é escrita de forma síncrona para o LUN externo e ONTAP. As duas cópias LUN são mantidas em sincronia dessa maneira até que o administrador execute uma transição completa que libera o LUN estrangeiro e não replica mais gravações.

O FLI foi projetado para funcionar com FC, mas se houver um desejo de mudar para iSCSI, o LUN migrado pode ser facilmente remapeado como um iSCSI LUN após a conclusão da migração.

Entre as caraterísticas da FLI está a deteção e ajuste automáticos do alinhamento. Neste contexto, o termo alinhamento refere-se a uma partição em um dispositivo LUN. O desempenho ideal requer que a e/S seja alinhada a 4K blocos. Se uma partição for colocada em um deslocamento que não é um múltiplo de 4K, o desempenho sofre.

Há um segundo aspeto do alinhamento que não pode ser corrigido ajustando um deslocamento de partição - o tamanho do bloco do sistema de arquivos. Por exemplo, um sistema de arquivos ZFS geralmente tem um tamanho de bloco interno de 512 bytes. Outros clientes que usam AIX criaram ocasionalmente sistemas de arquivos jfs2 com um tamanho de bloco de 512 ou 1, 024 bytes. Embora o sistema de arquivos possa estar alinhado a um limite 4K, os arquivos criados dentro desse sistema de arquivos não são e o desempenho sofre.

O FLI não deve ser usado nessas circunstâncias. Embora os dados estejam acessíveis após a migração, o resultado são sistemas de arquivos com sérias limitações de desempenho. Como princípio geral, qualquer sistema de arquivos que suporte uma carga de trabalho de substituição aleatória no ONTAP deve usar um tamanho de bloco 4K. Isso é principalmente aplicável a workloads como arquivos de dados de banco de dados e implantações de VDI. O tamanho do bloco pode ser identificado usando os comandos relevantes do sistema operacional do host.

Por exemplo, no AIX, o tamanho do bloco pode ser visualizado com `lsfs -q`o . Com Linux, `xfs_info` e `tune2fs` pode ser usado para `xfs` e `ext3/ext4`, respetivamente. `zfs`Com , o comando é `zdb -C`.

O parâmetro que controla o tamanho do bloco é e geralmente o padrão é `ashift` 9, o que significa 2 9, ou 512 bytes. Para um desempenho ideal, o `ashift` valor deve ser 12 (2 12 4K). Esse valor é definido no momento em que o zpool é criado e não pode ser alterado, o que significa que os zpools de dados com `ashift` outro que não o 12 devem ser migrados copiando dados para um zpool recém-criado.

O Oracle ASM não tem um tamanho de bloco fundamental. O único requisito é que a partição na qual o disco ASM é construído deve estar alinhada corretamente.

Ferramenta de transição de 7 modos

A 7-Mode Transition Tool (7MTT) é um utilitário de automação usado para migrar grandes configurações de 7 modos para o ONTAP. A maioria dos clientes de bancos de dados encontra outros métodos mais fáceis, em parte porque eles geralmente migram seus ambientes de banco de dados por banco de dados, em vez de realocar todo o espaço físico do storage. Além disso, os bancos de dados costumam fazer parte apenas de um ambiente de storage maior. Portanto, os bancos de dados geralmente são migrados individualmente e, em seguida, o ambiente restante pode ser movido com 7MTT.

Há um número pequeno, mas significativo de clientes que têm sistemas de storage dedicados a ambientes de banco de dados complicados. Esses ambientes podem conter muitos volumes, snapshots e vários detalhes de configuração, como permissões de exportação, grupos de iniciadores LUN, permissões de usuário e configuração do Lightweight Directory Access Protocol. Nesses casos, as habilidades de automação do 7MTT podem simplificar uma migração.

7MTT pode operar em um de dois modos:

- **Transição baseada em cópia (CBT).** O 7MTT com CBT configura volumes SnapMirror de um sistema de 7 modos existente no novo ambiente. Depois que os dados estiverem sincronizados, o 7MTT orquestrará o processo de transição.
- **Transição livre de cópia (CFT).** O 7MTT com CFT é baseado na conversão no local de prateleiras de disco existentes de 7 modos. Nenhum dado é copiado e os compartimentos de disco existentes podem ser reutilizados. Preserva a proteção de dados e a configuração de eficiência de storage existentes.

A principal diferença entre essas duas opções é que a transição livre de cópias é uma abordagem de grande impactos na qual todos os compartimentos de disco conetados ao par de HA de 7 modos original devem ser relocados para o novo ambiente. Não há opção de mover um subconjunto de prateleiras. A abordagem baseada em cópia permite que os volumes selecionados sejam movidos. Também há potencialmente uma janela de transição mais longa com transição livre de cópias por causa do laço necessário para reciclar compartimentos de disco e converter metadados. Com base na experiência de campo, a NetApp recomenda permitir 1 hora para realocação e desativação das gavetas de disco e entre 15 minutos e 2 horas para conversão de metadados.

Migração de arquivos de dados

Datafiles Oracle individuais podem ser movidos com um único comando.

Por exemplo, o comando a seguir move o arquivo de dados IOPST.dbf do sistema de arquivos para o sistema de `/oradata3` arquivos `/oradata2` .

```
SQL> alter database move datafile  '/oradata2/NTAP/IOPS002.dbf' to
'/oradata3/NTAP/IOPS002.dbf';
Database altered.
```

Mover um arquivo de dados com esse método pode ser lento, mas normalmente não deve produzir e/S suficiente para interferir com as cargas de trabalho diárias do banco de dados. Em contraste, a migração via rebalanceamento do ASM pode ser executada muito mais rápido, mas à custa de diminuir a velocidade do banco de dados geral enquanto os dados estão sendo movidos.

O tempo necessário para mover arquivos de dados pode ser facilmente medido criando um arquivo de dados de teste e, em seguida, movendo-o. O tempo decorrido para a operação é gravado nos dados da sessão:

```
SQL> set linesize 300;
SQL> select elapsed_seconds||':'||message from v$session_longops;
ELAPSED_SECONDS||':'||MESSAGE
------------------------------------------------------------------------
---------------
351:Online data file move: data file 8: 22548578304 out of 22548578304
bytes done
SQL> select bytes / 1024 / 1024 /1024 as GB from dba_data_files where
FILE_ID = 8;
        GB
----------
        21
```

Neste exemplo, o arquivo que foi movido foi o arquivo de dados 8, que era de 21GB MB de tamanho e exigiu cerca de 6 minutos para migrar. O tempo necessário obviamente depende dos recursos do sistema de storage, da rede de armazenamento e da atividade geral do banco de dados que ocorre no momento da migração.

Registrar envio

O objetivo de uma migração usando o envio de log é criar uma cópia dos arquivos de dados originais em um novo local e, em seguida, estabelecer um método de envio de alterações no novo ambiente.

Uma vez estabelecido, o envio de log e a repetição podem ser automatizados para manter o banco de dados de réplica em grande parte em sincronia com a fonte. Por exemplo, um trabalho cron pode ser programado para (a) copiar os logs mais recentes para o novo local e (b) reproduzi-los a cada 15 minutos. Isso proporciona uma interrupção mínima no momento da transição, porque não mais de 15 minutos de logs de arquivo devem ser reproduzidos.

O procedimento mostrado abaixo também é essencialmente uma operação clone do banco de dados. A lógica mostrada é semelhante ao mecanismo no NetApp SnapManager para Oracle (SMO) e no plug-in NetApp SnapCenter Oracle. Alguns clientes usaram o procedimento mostrado em scripts ou fluxos de trabalho WFA para operações de clonagem personalizadas. Embora esse procedimento seja mais manual do que o uso de SMO ou SnapCenter, ele ainda é facilmente programado e as APIs de gerenciamento de dados no ONTAP simplificam ainda mais o processo.

Envio de log - sistema de arquivos para sistema de arquivos

Este exemplo demonstra a migração de um banco de dados chamado WAFFLE de um sistema de arquivos comum para outro sistema de arquivos comum localizado em um servidor diferente. Ele também ilustra o uso do SnapMirror para fazer uma cópia rápida dos arquivos de dados, mas isso não é parte integrante do procedimento geral.

Criar cópia de segurança da base de dados

O primeiro passo é criar um backup de banco de dados. Especificamente, este procedimento requer um conjunto de arquivos de dados que podem ser usados para a repetição do log de arquivamento.

Ambiente

Neste exemplo, o banco de dados de origem está em um sistema ONTAP. O método mais simples para criar um backup de um banco de dados é usando um snapshot. O banco de dados é colocado no modo hot backup por alguns segundos enquanto uma `snapshot create` operação é executada no volume que hospeda os arquivos de dados.

```
SQL> alter database begin backup;
Database altered.
```

```
Cluster01::*> snapshot create -vserver vserver1 -volume jfsc1_oradata
hotbackup
Cluster01::*>
```

```
SQL> alter database end backup;
Database altered.
```

O resultado é um instantâneo no disco chamado `hotbackup` que contém uma imagem dos arquivos de dados enquanto no modo de backup automático. Quando combinados com os logs de arquivo apropriados para tornar os arquivos de dados consistentes, os dados nesse snapshot podem ser usados como base de uma restauração ou um clone. Neste caso, ele é replicado para o novo servidor.

Restauração de um novo ambiente

Agora, o backup precisa ser restaurado no novo ambiente. Isso pode ser feito de várias maneiras, incluindo Oracle RMAN, restauração de um aplicativo de backup como NetBackup, ou uma operação de cópia simples de arquivos de dados que foram colocados no modo de backup ativo.

Neste exemplo, o SnapMirror é usado para replicar o hotbackup instantâneo para um novo local.

1. Crie um novo volume para receber os dados instantâneos. Inicialize o espelhamento de `jfsc1_oradata` para `vol_oradata`.

   ```
   Cluster01::*> volume create -vserver vserver1 -volume vol_oradata
   -aggregate data_01 -size 20g -state online -type DP -snapshot-policy
   none -policy jfsc3
   [Job 833] Job succeeded: Successful
   ```

```
Cluster01::*> snapmirror initialize -source-path vserver1:jfsc1_oradata
-destination-path vserver1:vol_oradata
Operation is queued: snapmirror initialize of destination
"vserver1:vol_oradata".
Cluster01::*> volume mount -vserver vserver1 -volume vol_oradata
-junction-path /vol_oradata
Cluster01::*>
```

2. Depois que o estado é definido pelo SnapMirror, indicando que a sincronização está concluída, atualize o espelho com base especificamente no instantâneo desejado.

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields state
source-path             destination-path        state
----------------------- ----------------------- ------------
vserver1:jfsc1_oradata vserver1:vol_oradata SnapMirrored
```

```
Cluster01::*> snapmirror update -destination-path vserver1:vol_oradata
-source-snapshot hotbackup
Operation is queued: snapmirror update of destination
"vserver1:vol_oradata".
```

3. A sincronização bem-sucedida pode ser verificada visualizando o `newest-snapshot` campo no volume do espelho.

```
Cluster01::*> snapmirror show -destination-path vserver1:vol_oradata
-fields newest-snapshot
source-path             destination-path        newest-snapshot
----------------------- ----------------------- ---------------
vserver1:jfsc1_oradata vserver1:vol_oradata hotbackup
```

4. O espelho pode então ser quebrado.

```
Cluster01::> snapmirror break -destination-path vserver1:vol_oradata
Operation succeeded: snapmirror break for destination
"vserver1:vol_oradata".
Cluster01::>
```

5. Monte o novo sistema de arquivos. Com sistemas de arquivos baseados em blocos, os procedimentos precisos variam de acordo com a LVM em uso. O zoneamento FC ou as conexões iSCSI devem ser configuradas. Depois que a conetividade com os LUNs for estabelecida, comandos como o Linux `pvscan` podem ser necessários para descobrir quais grupos de volume ou LUNs precisam ser configurados

corretamente para serem descobertos pelo ASM.

Neste exemplo, um sistema de arquivos NFS simples é usado. Este sistema de arquivos pode ser montado diretamente.

```
fas8060-nfs1:/vol_oradata        19922944   1639360   18283584   9%
/oradata
fas8060-nfs1:/vol_logs            9961472       128    9961344   1%
/logs
```

Criar modelo de criação de controlfile

Em seguida, você deve criar um modelo de controlfile. O `backup controlfile to trace` comando cria comandos de texto para recriar um controlfile. Esta função pode ser útil para restaurar um banco de dados de backup em algumas circunstâncias, e é frequentemente usada com scripts que executam tarefas como clonagem de banco de dados.

1. A saída do comando a seguir é usada para recriar os controlfiles para o banco de dados migrado.

```
SQL> alter database backup controlfile to trace as '/tmp/waffle.ctrl';
Database altered.
```

2. Depois que os arquivos de controle tiverem sido criados, copie o arquivo para o novo servidor.

```
[oracle@jfsc3 tmp]$ scp oracle@jfsc1:/tmp/waffle.ctrl /tmp/
oracle@jfsc1's password:
waffle.ctrl                                           100% 5199
5.1KB/s   00:00
```

Ficheiro de parâmetros de cópia de segurança

Um arquivo de parâmetro também é necessário no novo ambiente. O método mais simples é criar um pfile a partir do spfile ou pfile atual. Neste exemplo, o banco de dados de origem está usando um spfile.

```
SQL> create pfile='/tmp/waffle.tmp.pfile' from spfile;
File created.
```

Crie uma entrada oratab

A criação de uma entrada oratab é necessária para o bom funcionamento de utilitários como oraenv. Para criar uma entrada oratab, execute a seguinte etapa.

```
WAFFLE:/orabin/product/12.1.0/dbhome_1:N
```

Prepare a estrutura do diretório

Se os diretórios necessários ainda não estavam presentes, você deve criá-los ou o procedimento de inicialização do banco de dados falhar. Para preparar a estrutura de diretórios, preencha os seguintes requisitos mínimos.

```
[oracle@jfsc3 ~]$ . oraenv
ORACLE_SID = [oracle] ? WAFFLE
The Oracle base has been set to /orabin
[oracle@jfsc3 ~]$ cd $ORACLE_BASE
[oracle@jfsc3 orabin]$ cd admin
[oracle@jfsc3 admin]$ mkdir WAFFLE
[oracle@jfsc3 admin]$ cd WAFFLE
[oracle@jfsc3 WAFFLE]$ mkdir adump dpdump pfile scripts xdb_wallet
```

Atualizações do arquivo de parâmetros

1. Para copiar o arquivo de parâmetro para o novo servidor, execute os seguintes comandos. O local padrão é o `$ORACLE_HOME/dbs` diretório. Neste caso, o pfile pode ser colocado em qualquer lugar. Ele só está sendo usado como um passo intermediário no processo de migração.

```
[oracle@jfsc3 admin]$ scp oracle@jfsc1:/tmp/waffle.tmp.pfile
$ORACLE_HOME/dbs/waffle.tmp.pfile
oracle@jfsc1's password:
waffle.pfile                                          100%  916
0.9KB/s   00:00
```

1. Edite o arquivo conforme necessário. Por exemplo, se a localização do log do arquivo foi alterada, o arquivo pfile deve ser alterado para refletir o novo local. Neste exemplo, apenas os controlfiles estão sendo relocados, em parte para distribuí-los entre os sistemas de arquivos de log e dados.

```
[root@jfsc1 tmp]# cat waffle.pfile
WAFFLE.__data_transfer_cache_size=0
WAFFLE.__db_cache_size=507510784
WAFFLE.__java_pool_size=4194304
WAFFLE.__large_pool_size=20971520
WAFFLE.__oracle_base='/orabin'#ORACLE_BASE set from environment
WAFFLE.__pga_aggregate_target=268435456
WAFFLE.__sga_target=805306368
WAFFLE.__shared_io_pool_size=29360128
WAFFLE.__shared_pool_size=234881024
WAFFLE.__streams_pool_size=0
*.audit_file_dest='/orabin/admin/WAFFLE/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='/oradata//WAFFLE/control01.ctl','/oradata//WAFFLE/contr
ol02.ctl'
*.control_files='/oradata/WAFFLE/control01.ctl','/logs/WAFFLE/control02.
ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='WAFFLE'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=WAFFLEXDB)'
*.log_archive_dest_1='LOCATION=/logs/WAFFLE/arch'
*.log_archive_format='%t_%s_%r.dbf'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

2. Depois que as edições estiverem concluídas, crie um spfile baseado nesse pfile.

```
SQL> create spfile from pfile='waffle.tmp.pfile';
File created.
```

Recrie controlfiles

Em uma etapa anterior, a saída do `backup controlfile to trace` foi copiada para o novo servidor. A parte específica da saída necessária é o `controlfile recreation` comando. Esta informação pode ser encontrada no ficheiro na secção marcada `Set #1. NORESETLOGS`. Começa com a linha `create controlfile reuse database` e deve incluir a palavra `noresetlogs`. Termina com o caractere ponto e vírgula (;).

1. Neste procedimento de exemplo, o arquivo lê o seguinte.

```
CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
    MAXLOGFILES 16
    MAXLOGMEMBERS 3
    MAXDATAFILES 100
    MAXINSTANCES 8
    MAXLOGHISTORY 292
LOGFILE
  GROUP 1 '/logs/WAFFLE/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 2 '/logs/WAFFLE/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 3 '/logs/WAFFLE/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
  '/oradata/WAFFLE/system01.dbf',
  '/oradata/WAFFLE/sysaux01.dbf',
  '/oradata/WAFFLE/undotbs01.dbf',
  '/oradata/WAFFLE/users01.dbf'
CHARACTER SET WE8MSWIN1252
;
```

2. Edite este script como desejado para refletir a nova localização dos vários arquivos. Por exemplo, certos arquivos de dados conhecidos por oferecer suporte a e/S alta podem ser redirecionados para um sistema de arquivos em uma camada de storage de alto desempenho. Em outros casos, as alterações podem ser puramente por razões de administrador, como isolar os arquivos de dados de um determinado PDB em volumes dedicados.
3. Neste exemplo, a `DATAFILE` estrofe permanece inalterada, mas os logs de refazer são movidos para um novo local em `/redo` vez de compartilhar espaço com logs de arquivo no `/logs`.

```
CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
    MAXLOGFILES 16
    MAXLOGMEMBERS 3
    MAXDATAFILES 100
    MAXINSTANCES 8
    MAXLOGHISTORY 292
LOGFILE
  GROUP 1 '/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 2 '/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
  GROUP 3 '/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
  '/oradata/WAFFLE/system01.dbf',
  '/oradata/WAFFLE/sysaux01.dbf',
  '/oradata/WAFFLE/undotbs01.dbf',
  '/oradata/WAFFLE/users01.dbf'
CHARACTER SET WE8MSWIN1252
;
```

```
SQL> startup nomount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             331353200 bytes
Database Buffers          465567744 bytes
Redo Buffers                5455872 bytes
SQL> CREATE CONTROLFILE REUSE DATABASE "WAFFLE" NORESETLOGS  ARCHIVELOG
  2      MAXLOGFILES 16
  3      MAXLOGMEMBERS 3
  4      MAXDATAFILES 100
  5      MAXINSTANCES 8
  6      MAXLOGHISTORY 292
  7  LOGFILE
  8    GROUP 1 '/redo/redo01.log'  SIZE 50M BLOCKSIZE 512,
  9    GROUP 2 '/redo/redo02.log'  SIZE 50M BLOCKSIZE 512,
 10    GROUP 3 '/redo/redo03.log'  SIZE 50M BLOCKSIZE 512
 11  -- STANDBY LOGFILE
 12  DATAFILE
 13    '/oradata/WAFFLE/system01.dbf',
 14    '/oradata/WAFFLE/sysaux01.dbf',
 15    '/oradata/WAFFLE/undotbs01.dbf',
 16    '/oradata/WAFFLE/users01.dbf'
 17  CHARACTER SET WE8MSWIN1252
 18  ;
Control file created.
SQL>
```

Se algum arquivo estiver perdido ou os parâmetros estiverem mal configurados, são gerados erros que indicam o que deve ser corrigido. O banco de dados está montado, mas ainda não está aberto e não pode ser aberto porque os arquivos de dados em uso ainda estão marcados como estando no modo hot backup. Os logs de arquivamento devem primeiro ser aplicados para tornar o banco de dados consistente.

Replicação inicial do log

Pelo menos uma operação de resposta de log é necessária para tornar os arquivos de dados consistentes. Muitas opções estão disponíveis para reproduzir logs. Em alguns casos, o local do log do arquivo original no servidor original pode ser compartilhado por NFS e a resposta do log pode ser feita diretamente. Em outros casos, os logs do arquivo devem ser copiados.

Por exemplo, uma operação simples `scp` pode copiar todos os logs atuais do servidor de origem para o servidor de migração:

```
[oracle@jfsc3 arch]$ scp jfsc1:/logs/WAFFLE/arch/* ./
oracle@jfsc1's password:
1_22_912662036.dbf                                          100%   47MB
47.0MB/s   00:01
1_23_912662036.dbf                                          100%   40MB
40.4MB/s   00:00
1_24_912662036.dbf                                          100%   45MB
45.4MB/s   00:00
1_25_912662036.dbf                                          100%   41MB
40.9MB/s   00:01
1_26_912662036.dbf                                          100%   39MB
39.4MB/s   00:00
1_27_912662036.dbf                                          100%   39MB
38.7MB/s   00:00
1_28_912662036.dbf                                          100%   40MB
40.1MB/s   00:01
1_29_912662036.dbf                                          100%   17MB
16.9MB/s   00:00
1_30_912662036.dbf                                          100%  636KB
636.0KB/s   00:00
```

Reprodução inicial do registo

Depois que os arquivos estão no local do log de arquivamento, eles podem ser reproduzidos emitindo o comando `recover database until cancel` seguido da resposta `AUTO` para reproduzir automaticamente todos os logs disponíveis.

```
SQL> recover database until cancel;
ORA-00279: change 382713 generated at 05/24/2016 09:00:54 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_23_912662036.dbf
ORA-00280: change 382713 for thread 1 is in sequence #23
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 405712 generated at 05/24/2016 15:01:05 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_24_912662036.dbf
ORA-00280: change 405712 for thread 1 is in sequence #24
ORA-00278: log file '/logs/WAFFLE/arch/1_23_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
ORA-00278: log file '/logs/WAFFLE/arch/1_30_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_31_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

A resposta final do log do arquivo relata um erro, mas isso é normal. O log indica que `sqlplus` estava procurando um arquivo de log específico e não o encontrou. A razão é, muito provavelmente, que o arquivo log ainda não existe.

Se o banco de dados de origem puder ser desligado antes de copiar logs de arquivo, esta etapa deve ser executada apenas uma vez. Os logs de arquivo são copiados e reproduzidos e, em seguida, o processo pode continuar diretamente para o processo de transição que replica os logs críticos de refazer.

Replicação e repetição de registos incrementais

Na maioria dos casos, a migração não é realizada imediatamente. Pode ser dias ou mesmo semanas antes que o processo de migração seja concluído, o que significa que os logs devem ser enviados continuamente para o banco de dados de réplica e reproduzidos. Portanto, quando a transição chega, os dados mínimos devem ser transferidos e reproduzidos.

Fazer isso pode ser script de várias maneiras, mas um dos métodos mais populares é usar rsync, um utilitário comum de replicação de arquivos. A maneira mais segura de usar este utilitário é configurá-lo como um daemon. Por exemplo, o `rsyncd.conf` arquivo a seguir mostra como criar um recurso `waffle.arch` chamado que é acessado com credenciais de usuário Oracle e é mapeado para `/logs/WAFFLE/arch`o . Mais importante ainda, o recurso é definido como somente leitura, o que permite que os dados de produção sejam lidos, mas não alterados.

```
[root@jfsc1 arch]# cat /etc/rsyncd.conf
[waffle.arch]
   uid=oracle
   gid=dba
   path=/logs/WAFFLE/arch
   read only = true
[root@jfsc1 arch]# rsync --daemon
```

O comando a seguir sincroniza o destino do log de arquivamento do novo servidor com o recurso rsync `waffle.arch` no servidor original. O `t` argumento em `rsync - potg` faz com que a lista de arquivos seja comparada com base no timestamp, e apenas novos arquivos são copiados. Este processo fornece uma atualização incremental do novo servidor. Esse comando também pode ser programado no cron para ser executado regularmente.

```
[oracle@jfsc3 arch]$ rsync -potg --stats --progress jfsc1::waffle.arch/*
/logs/WAFFLE/arch/
1_31_912662036.dbf
      650240 100%  124.02MB/s    0:00:00 (xfer#1, to-check=8/18)
1_32_912662036.dbf
     4873728 100%  110.67MB/s    0:00:00 (xfer#2, to-check=7/18)
1_33_912662036.dbf
     4088832 100%   50.64MB/s    0:00:00 (xfer#3, to-check=6/18)
1_34_912662036.dbf
     8196096 100%   54.66MB/s    0:00:00 (xfer#4, to-check=5/18)
1_35_912662036.dbf
    19376128 100%   57.75MB/s    0:00:00 (xfer#5, to-check=4/18)
1_36_912662036.dbf
       71680 100%  201.15kB/s    0:00:00 (xfer#6, to-check=3/18)
1_37_912662036.dbf
     1144320 100%    3.06MB/s    0:00:00 (xfer#7, to-check=2/18)
1_38_912662036.dbf
    35757568 100%   63.74MB/s    0:00:00 (xfer#8, to-check=1/18)
1_39_912662036.dbf
      984576 100%    1.63MB/s    0:00:00 (xfer#9, to-check=0/18)
Number of files: 18
Number of files transferred: 9
Total file size: 399653376 bytes
Total transferred file size: 75143168 bytes
Literal data: 75143168 bytes
Matched data: 0 bytes
File list size: 474
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 204
Total bytes received: 75153219
sent 204 bytes  received 75153219 bytes  150306846.00 bytes/sec
total size is 399653376  speedup is 5.32
```

Depois que os logs tiverem sido recebidos, eles devem ser reproduzidos novamente. Exemplos anteriores mostram o uso do sqlplus para executar manualmente `recover database until cancel` , um processo que pode ser facilmente automatizado. O exemplo mostrado aqui usa o script descrito em "Reproduzir Registos na base de dados". Os scripts aceitam um argumento que especifica o banco de dados que requer uma operação de repetição. Isso permite que o mesmo script seja usado em um esforço de migração multibanco de dados.

```
[oracle@jfsc3 logs]$ ./replay.logs.pl WAFFLE
ORACLE_SID = [WAFFLE] ? The Oracle base remains unchanged with value
/orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu May 26 10:47:16 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 713874 generated at 05/26/2016 04:26:43 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_31_912662036.dbf
ORA-00280: change 713874 for thread 1 is in sequence #31
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 814256 generated at 05/26/2016 04:52:30 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_32_912662036.dbf
ORA-00280: change 814256 for thread 1 is in sequence #32
ORA-00278: log file '/logs/WAFFLE/arch/1_31_912662036.dbf' no longer
needed for
this recovery
ORA-00279: change 814780 generated at 05/26/2016 04:53:04 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_33_912662036.dbf
ORA-00280: change 814780 for thread 1 is in sequence #33
ORA-00278: log file '/logs/WAFFLE/arch/1_32_912662036.dbf' no longer
needed for
this recovery
...
ORA-00279: change 1120099 generated at 05/26/2016 09:59:21 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_40_912662036.dbf
ORA-00280: change 1120099 for thread 1 is in sequence #40
ORA-00278: log file '/logs/WAFFLE/arch/1_39_912662036.dbf' no longer
needed for
this recovery
ORA-00308: cannot open archived log '/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

Redução

Quando você estiver pronto para cortar para o novo ambiente, você deve executar uma sincronização final que inclua Registros de arquivamento e os logs de refazer. Se a localização original do registo de refazer ainda não for conhecida, pode ser identificada da seguinte forma:

```
SQL> select member from v$logfile;
MEMBER
--------------------------------------------------------------------------------
------
/logs/WAFFLE/redo/redo01.log
/logs/WAFFLE/redo/redo02.log
/logs/WAFFLE/redo/redo03.log
```

1. Encerre o banco de dados de origem.
2. Execute uma sincronização final dos logs de arquivo no novo servidor com o método desejado.
3. Os logs de refazer de origem devem ser copiados para o novo servidor. Neste exemplo, os logs de refazer foram relocados para um novo diretório em `/redo`.

   ```
   [oracle@jfsc3 logs]$ scp jfsc1:/logs/WAFFLE/redo/* /redo/
   oracle@jfsc1's password:
   redo01.log
   100%   50MB  50.0MB/s   00:01
   redo02.log
   100%   50MB  50.0MB/s   00:00
   redo03.log
   100%   50MB  50.0MB/s   00:00
   ```

4. Neste estágio, o novo ambiente de banco de dados contém todos os arquivos necessários para trazê-lo para o mesmo estado exato da origem. Os registos de arquivo têm de ser reproduzidos uma última vez.

```
SQL> recover database until cancel;
ORA-00279: change 1120099 generated at 05/26/2016 09:59:21 needed for
thread 1
ORA-00289: suggestion : /logs/WAFFLE/arch/1_40_912662036.dbf
ORA-00280: change 1120099 for thread 1 is in sequence #40
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00308: cannot open archived log
'/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
ORA-00308: cannot open archived log
'/logs/WAFFLE/arch/1_40_912662036.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

5. Uma vez concluído, os logs de refazer devem ser reproduzidos novamente. Se a mensagem `Media recovery complete` for retornada, o processo será bem-sucedido e os bancos de dados serão sincronizados e poderão ser abertos.

```
SQL> recover database;
Media recovery complete.
SQL> alter database open;
Database altered.
```

Registo de envio - ASM para o sistema de ficheiros

Este exemplo demonstra o uso do Oracle RMAN para migrar um banco de dados. É muito semelhante ao exemplo anterior de sistema de arquivos para o envio de log do sistema de arquivos, mas os arquivos no ASM não são visíveis para o host. As únicas opções de migração de dados localizados em dispositivos ASM são a realocação do ASM LUN ou o Oracle RMAN para executar as operações de cópia.

Embora o RMAN seja um requisito para copiar arquivos do Oracle ASM, o uso do RMAN não se limita ao ASM. O RMAN pode ser usado para migrar de qualquer tipo de armazenamento para qualquer outro tipo.

Este exemplo mostra a realocação de um banco de dados chamado PANCAKE do armazenamento ASM para um sistema de arquivos regular localizado em um servidor diferente em caminhos `/oradata` e `/logs`.

Criar cópia de segurança da base de dados

O primeiro passo é criar um backup do banco de dados para ser migrado para um servidor alternativo. Como a fonte usa o Oracle ASM, o RMAN deve ser usado. Um simples backup RMAN pode ser executado da seguinte forma. Este método cria um backup marcado que pode ser facilmente identificado pelo RMAN mais tarde no procedimento.

O primeiro comando define o tipo de destino para o backup e o local a ser usado. O segundo inicia o backup dos arquivos de dados somente.

```
RMAN> configure channel device type disk format '/rman/pancake/%U';
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT   '/rman/pancake/%U';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT   '/rman/pancake/%U';
new RMAN configuration parameters are successfully stored
RMAN> backup database tag 'ONTAP_MIGRATION';
Starting backup at 24-MAY-16
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=251 device type=DISK
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001 name=+ASM0/PANCAKE/system01.dbf
input datafile file number=00002 name=+ASM0/PANCAKE/sysaux01.dbf
input datafile file number=00003 name=+ASM0/PANCAKE/undotbs101.dbf
input datafile file number=00004 name=+ASM0/PANCAKE/users01.dbf
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/1gr6c161_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:03
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/1hr6c164_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Ficheiro de controlo de cópia de segurança

Um ficheiro de controlo de cópia de segurança é necessário mais tarde no procedimento para a `duplicate database` operação.

```
RMAN> backup current controlfile format '/rman/pancake/ctrl.bkp';
Starting backup at 24-MAY-16
using channel ORA_DISK_1
channel ORA_DISK_1: starting full datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
channel ORA_DISK_1: starting piece 1 at 24-MAY-16
channel ORA_DISK_1: finished piece 1 at 24-MAY-16
piece handle=/rman/pancake/ctrl.bkp tag=TAG20160524T032651 comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Ficheiro de parâmetros de cópia de segurança

Um arquivo de parâmetro também é necessário no novo ambiente. O método mais simples é criar um pfile a partir do spfile ou pfile atual. Neste exemplo, o banco de dados de origem usa um spfile.

```
RMAN> create pfile='/rman/pancake/pfile' from spfile;
Statement processed
```

Script de renomeação do arquivo ASM

Vários locais de arquivo atualmente definidos nos controlfiles mudam quando o banco de dados é movido. O script a seguir cria um script RMAN para facilitar o processo. Este exemplo mostra um banco de dados com um número muito pequeno de arquivos de dados, mas normalmente os bancos de dados contêm centenas ou até mesmo milhares de arquivos de dados.

Este script pode ser encontrado em "ASM para conversão de nome de sistema de arquivos" e faz duas coisas.

Primeiro, ele cria um parâmetro para redefinir os locais de log refazer chamados `log_file_name_convert`. É essencialmente uma lista de campos alternados. O primeiro campo é a localização de um log de refazer atual e o segundo campo é a localização no novo servidor. O padrão é então repetido.

A segunda função é fornecer um modelo para renomeação de arquivos de dados. O script percorre os arquivos de dados, puxa as informações de nome e número do arquivo e formata-as como um script RMAN. Em seguida, ele faz o mesmo com os arquivos temporários. O resultado é um script rman simples que pode ser editado como desejado para garantir que os arquivos sejam restaurados para o local desejado.

```
SQL> @/rman/mk.rename.scripts.sql
Parameters for log file conversion:
*.log_file_name_convert = '+ASM0/PANCAKE/redo01.log',
'/NEW_PATH/redo01.log','+ASM0/PANCAKE/redo02.log',
'/NEW_PATH/redo02.log','+ASM0/PANCAKE/redo03.log', '/NEW_PATH/redo03.log'
rman duplication script:
run
{
set newname for datafile 1 to '+ASM0/PANCAKE/system01.dbf';
set newname for datafile 2 to '+ASM0/PANCAKE/sysaux01.dbf';
set newname for datafile 3 to '+ASM0/PANCAKE/undotbs101.dbf';
set newname for datafile 4 to '+ASM0/PANCAKE/users01.dbf';
set newname for tempfile 1 to '+ASM0/PANCAKE/temp01.dbf';
duplicate target database for standby backup location INSERT_PATH_HERE;
}
PL/SQL procedure successfully completed.
```

Capture a saída desta tela. O `log_file_name_convert` parâmetro é colocado no arquivo pfile como descrito abaixo. O nome do arquivo de dados RMAN e o script duplicado devem ser editados de acordo para colocar os arquivos de dados nos locais desejados. Neste exemplo, todos eles são colocados `/oradata/pancake` em .

```
run
{
set newname for datafile 1 to '/oradata/pancake/pancake.dbf';
set newname for datafile 2 to '/oradata/pancake/sysaux.dbf';
set newname for datafile 3 to '/oradata/pancake/undotbs1.dbf';
set newname for datafile 4 to '/oradata/pancake/users.dbf';
set newname for tempfile 1 to '/oradata/pancake/temp.dbf';
duplicate target database for standby backup location '/rman/pancake';
}
```

Prepare a estrutura do diretório

Os scripts estão quase prontos para serem executados, mas primeiro a estrutura de diretórios deve estar no lugar. Se os diretórios necessários ainda não estiverem presentes, eles devem ser criados ou o procedimento de inicialização do banco de dados falha. O exemplo abaixo reflete os requisitos mínimos.

```
[oracle@jfsc2 ~]$ mkdir /oradata/pancake
[oracle@jfsc2 ~]$ mkdir /logs/pancake
[oracle@jfsc2 ~]$ cd /orabin/admin
[oracle@jfsc2 admin]$ mkdir PANCAKE
[oracle@jfsc2 admin]$ cd PANCAKE
[oracle@jfsc2 PANCAKE]$ mkdir adump dpdump pfile scripts xdb_wallet
```

Crie uma entrada oratab

O comando a seguir é necessário para que utilitários como oraenv funcionem corretamente.

```
PANCAKE:/orabin/product/12.1.0/dbhome_1:N
```

Atualizações de parâmetros

O arquivo pfile salvo deve ser atualizado para refletir quaisquer alterações de caminho no novo servidor. As alterações no caminho do arquivo de dados são alteradas pelo script de duplicação RMAN, e quase todos os bancos de dados exigem alterações nos `control_files` parâmetros e. `log_archive_dest` Também pode haver locais de arquivo de auditoria que devem ser alterados e parâmetros como `db_create_file_dest` podem não ser relevantes fora do ASM. Um DBA experiente deve analisar cuidadosamente as alterações propostas antes de prosseguir.

Neste exemplo, as alterações de chave são as localizações do arquivo de controle, o destino do arquivo de log e a adição do `log_file_name_convert` parâmetro.

```
PANCAKE.__data_transfer_cache_size=0
PANCAKE.__db_cache_size=545259520
PANCAKE.__java_pool_size=4194304
PANCAKE.__large_pool_size=25165824
PANCAKE.__oracle_base='/orabin'#ORACLE_BASE set from environment
PANCAKE.__pga_aggregate_target=268435456
PANCAKE.__sga_target=805306368
PANCAKE.__shared_io_pool_size=29360128
PANCAKE.__shared_pool_size=192937984
PANCAKE.__streams_pool_size=0
*.audit_file_dest='/orabin/admin/PANCAKE/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='+ASM0/PANCAKE/control01.ctl','+ASM0/PANCAKE/control02.ctl
'
*.control_files='/oradata/pancake/control01.ctl','/logs/pancake/control02.
ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='PANCAKE'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=PANCAKEXDB)'
*.log_archive_dest_1='LOCATION=+ASM1'
*.log_archive_dest_1='LOCATION=/logs/pancake'
*.log_archive_format='%t_%s_%r.dbf'
'/logs/path/redo02.log'
*.log_file_name_convert = '+ASM0/PANCAKE/redo01.log',
'/logs/pancake/redo01.log', '+ASM0/PANCAKE/redo02.log',
'/logs/pancake/redo02.log', '+ASM0/PANCAKE/redo03.log',
'/logs/pancake/redo03.log'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

Depois que os novos parâmetros são confirmados, os parâmetros devem ser colocados em vigor. Existem várias opções, mas a maioria dos clientes cria um spfile baseado no pfile de texto.

```
bash-4.1$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0 Production on Fri Jan 8 11:17:40 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> create spfile from pfile='/rman/pancake/pfile';
File created.
```

Nomunt de arranque

A etapa final antes de replicar o banco de dados é abrir os processos do banco de dados, mas não montar os arquivos. Nesta etapa, os problemas com o spfile podem se tornar evidentes. Se o `startup nomount` comando falhar por causa de um erro de parâmetro, é simples desligar, corrigir o modelo pfile, recarregá-lo como um spfile e tentar novamente.

```
SQL> startup nomount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             373296240 bytes
Database Buffers          423624704 bytes
Redo Buffers                5455872 bytes
```

Duplique o banco de dados

Restaurar o backup RMAN anterior para o novo local consome mais tempo do que outras etapas deste processo. O banco de dados deve ser duplicado sem uma alteração no ID do banco de dados (DBID) ou redefinir os logs. Isso impede que os logs sejam aplicados, que é uma etapa necessária para sincronizar totalmente as cópias.

Conete-se ao banco de dados com RMAN como aux e emita o comando duplicar banco de dados usando o script criado em uma etapa anterior.

```
[oracle@jfsc2 pancake]$ rman auxiliary /
Recovery Manager: Release 12.1.0.2.0 - Production on Tue May 24 03:04:56
2016
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to auxiliary database: PANCAKE (not mounted)
RMAN> run
2> {
3> set newname for datafile 1 to '/oradata/pancake/pancake.dbf';
4> set newname for datafile 2 to '/oradata/pancake/sysaux.dbf';
5> set newname for datafile 3 to '/oradata/pancake/undotbs1.dbf';
6> set newname for datafile 4 to '/oradata/pancake/users.dbf';
7> set newname for tempfile 1 to '/oradata/pancake/temp.dbf';
```

```
8> duplicate target database for standby backup location '/rman/pancake';
9> }
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
Starting Duplicate Db at 24-MAY-16
contents of Memory Script:
{
   restore clone standby controlfile from  '/rman/pancake/ctrl.bkp';
}
executing Memory Script
Starting restore at 24-MAY-16
allocated channel: ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: SID=243 device type=DISK
channel ORA_AUX_DISK_1: restoring control file
channel ORA_AUX_DISK_1: restore complete, elapsed time: 00:00:01
output file name=/oradata/pancake/control01.ctl
output file name=/logs/pancake/control02.ctl
Finished restore at 24-MAY-16
contents of Memory Script:
{
   sql clone 'alter database mount standby database';
}
executing Memory Script
sql statement: alter database mount standby database
released channel: ORA_AUX_DISK_1
allocated channel: ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: SID=243 device type=DISK
contents of Memory Script:
{
   set newname for tempfile  1 to
 "/oradata/pancake/temp.dbf";
   switch clone tempfile all;
   set newname for datafile  1 to
 "/oradata/pancake/pancake.dbf";
   set newname for datafile  2 to
 "/oradata/pancake/sysaux.dbf";
   set newname for datafile  3 to
 "/oradata/pancake/undotbs1.dbf";
   set newname for datafile  4 to
 "/oradata/pancake/users.dbf";
   restore
   clone database
   ;
```

```
}
executing Memory Script
executing command: SET NEWNAME
renamed tempfile 1 to /oradata/pancake/temp.dbf in control file
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
executing command: SET NEWNAME
Starting restore at 24-MAY-16
using channel ORA_AUX_DISK_1
channel ORA_AUX_DISK_1: starting datafile backup set restore
channel ORA_AUX_DISK_1: specifying datafile(s) to restore from backup set
channel ORA_AUX_DISK_1: restoring datafile 00001 to
/oradata/pancake/pancake.dbf
channel ORA_AUX_DISK_1: restoring datafile 00002 to
/oradata/pancake/sysaux.dbf
channel ORA_AUX_DISK_1: restoring datafile 00003 to
/oradata/pancake/undotbs1.dbf
channel ORA_AUX_DISK_1: restoring datafile 00004 to
/oradata/pancake/users.dbf
channel ORA_AUX_DISK_1: reading from backup piece
/rman/pancake/1gr6c161_1_1
channel ORA_AUX_DISK_1: piece handle=/rman/pancake/1gr6c161_1_1
tag=ONTAP_MIGRATION
channel ORA_AUX_DISK_1: restored backup piece 1
channel ORA_AUX_DISK_1: restore complete, elapsed time: 00:00:07
Finished restore at 24-MAY-16
contents of Memory Script:
{
   switch clone datafile all;
}
executing Memory Script
datafile 1 switched to datafile copy
input datafile copy RECID=5 STAMP=912655725 file
name=/oradata/pancake/pancake.dbf
datafile 2 switched to datafile copy
input datafile copy RECID=6 STAMP=912655725 file
name=/oradata/pancake/sysaux.dbf
datafile 3 switched to datafile copy
input datafile copy RECID=7 STAMP=912655725 file
name=/oradata/pancake/undotbs1.dbf
datafile 4 switched to datafile copy
input datafile copy RECID=8 STAMP=912655725 file
name=/oradata/pancake/users.dbf
Finished Duplicate Db at 24-MAY-16
```

Replicação inicial do log

Agora você deve enviar as alterações do banco de dados de origem para um novo local. Fazer isso pode exigir uma combinação de etapas. O método mais simples seria fazer com que o RMAN no banco de dados de origem escrevesse logs de arquivo em uma conexão de rede compartilhada. Se um local compartilhado não estiver disponível, um método alternativo é usar o RMAN para gravar em um sistema de arquivos local e, em seguida, usar RCP ou rsync para copiar os arquivos.

Neste exemplo, o `/rman` diretório é um compartilhamento NFS disponível para o banco de dados original e migrado.

Uma questão importante aqui é a `disk format` cláusula. O formato de disco do backup é `%h_%e_%a.dbf`, o que significa que você deve usar o formato de número de thread, número de sequência e ID de ativação para o banco de dados. Embora as letras sejam diferentes, isso corresponde ao `log_archive_format='%t_%s_%r.dbf` parâmetro no pfile. Este parâmetro também especifica Registros de arquivo no formato de número de thread, número de sequência e ID de ativação. O resultado final é que os backups dos arquivos de log na fonte usam uma convenção de nomenclatura esperada pelo banco de dados. Isso torna operações como `recover database` muito mais simples porque sqlplus antecipa corretamente os nomes dos logs do arquivo a serem reproduzidos.

```
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/arch/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
released channel: ORA_DISK_1
RMAN> backup as copy archivelog from time 'sysdate-2';
Starting backup at 24-MAY-16
current log archived
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=373 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=70 STAMP=912658508
output file name=/rman/pancake/logship/1_54_912576125.dbf RECID=123
STAMP=912659482
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=41 RECID=29 STAMP=912654101
output file name=/rman/pancake/logship/1_41_912576125.dbf RECID=124
STAMP=912659483
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=33 STAMP=912654688
output file name=/rman/pancake/logship/1_45_912576125.dbf RECID=152
STAMP=912659514
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=36 STAMP=912654809
output file name=/rman/pancake/logship/1_47_912576125.dbf RECID=153
STAMP=912659515
channel ORA_DISK_1: archived log copy complete, elapsed time: 00:00:01
Finished backup at 24-MAY-16
```

Reprodução inicial do registo

Depois que os arquivos estão no local do log de arquivamento, eles podem ser reproduzidos emitindo o comando `recover database until cancel` seguido da resposta `AUTO` para reproduzir automaticamente todos os logs disponíveis. O arquivo de parâmetros está direcionando Registros de arquivo para `/logs/archive`, mas isso não corresponde ao local onde o RMAN foi usado para salvar logs. O local pode ser temporariamente redirecionado da seguinte forma antes de recuperar o banco de dados.

```
SQL> alter system set log_archive_dest_1='LOCATION=/rman/pancake/logship'
scope=memory;
System altered.
SQL> recover standby database until cancel;
ORA-00279: change 560224 generated at 05/24/2016 03:25:53 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_49_912576125.dbf
ORA-00280: change 560224 for thread 1 is in sequence #49
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
AUTO
ORA-00279: change 560353 generated at 05/24/2016 03:29:17 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_50_912576125.dbf
ORA-00280: change 560353 for thread 1 is in sequence #50
ORA-00278: log file '/rman/pancake/logship/1_49_912576125.dbf' no longer
needed
for this recovery
...
ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
ORA-00278: log file '/rman/pancake/logship/1_53_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_54_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
```

A resposta final do log do arquivo relata um erro, mas isso é normal. O erro indica que sqlplus estava procurando um arquivo de log específico e não o encontrou. A razão é mais provável que o arquivo log ainda não existe.

Se o banco de dados de origem puder ser desligado antes de copiar logs de arquivo, esta etapa deve ser executada apenas uma vez. Os logs de arquivo são copiados e reproduzidos e, em seguida, o processo pode continuar diretamente para o processo de transição que replica os logs críticos de refazer.

Replicação e repetição de registos incrementais

Na maioria dos casos, a migração não é realizada imediatamente. Pode ser dias ou mesmo semanas antes que o processo de migração seja concluído, o que significa que os logs devem ser continuamente enviados para o banco de dados de réplica e reproduzidos. Isso garante que os dados mínimos devem ser transferidos e reproduzidos quando a transição chegar.

Este processo pode ser facilmente programado. Por exemplo, o comando a seguir pode ser agendado no banco de dados original para garantir que o local usado para o envio de logs seja atualizado continuamente.

```
[oracle@jfsc1 pancake]$ cat copylogs.rman
configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
backup as copy archivelog from time 'sysdate-2';
```

```
[oracle@jfsc1 pancake]$ rman target / cmdfile=copylogs.rman
Recovery Manager: Release 12.1.0.2.0 - Production on Tue May 24 04:36:19
2016
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to target database: PANCAKE (DBID=3574534589)
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
2> backup as copy archivelog from time 'sysdate-2';
3>
4>
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
Starting backup at 24-MAY-16
current log archived
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=369 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=123 STAMP=912659482
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:22
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_54_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=41 RECID=124 STAMP=912659483
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:23
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_41_912576125.dbf
continuing other job steps, job failed will not be re-run
...
channel ORA_DISK_1: starting archived log copy
```

```
input archived log thread=1 sequence=45 RECID=152 STAMP=912659514
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:55
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_45_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=153 STAMP=912659515
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03009: failure of backup command on ORA_DISK_1 channel at 05/24/2016
04:36:57
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_47_912576125.dbf
Recovery Manager complete.
```

Depois que os logs tiverem sido recebidos, eles devem ser reproduzidos novamente. Exemplos anteriores mostraram o uso do sqlplus para executar manualmente `recover database until cancel` , que pode ser facilmente automatizado. O exemplo mostrado aqui usa o script descrito em "Reproduzir Registos na base de dados em espera". O script aceita um argumento que especifica o banco de dados que requer uma operação de repetição. Esse processo permite que o mesmo script seja usado em um esforço de migração de multibanco de dados.

```
[root@jfsc2 pancake]# ./replaylogs.pl PANCAKE
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Tue May 24 04:47:10 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 560591 generated at 05/24/2016 03:33:56 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_54_912576125.dbf
ORA-00280: change 560591 for thread 1 is in sequence #54
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 562219 generated at 05/24/2016 04:15:08 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_55_912576125.dbf
ORA-00280: change 562219 for thread 1 is in sequence #55
ORA-00278: log file '/rman/pancake/logship/1_54_912576125.dbf' no longer
needed for this recovery
ORA-00279: change 562370 generated at 05/24/2016 04:19:18 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_56_912576125.dbf
ORA-00280: change 562370 for thread 1 is in sequence #56
ORA-00278: log file '/rman/pancake/logship/1_55_912576125.dbf' no longer
needed for this recovery
...
ORA-00279: change 563137 generated at 05/24/2016 04:36:20 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_65_912576125.dbf
ORA-00280: change 563137 for thread 1 is in sequence #65
ORA-00278: log file '/rman/pancake/logship/1_64_912576125.dbf' no longer
needed for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_65_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

Redução

Quando estiver pronto para ser cortado para o novo ambiente, você deve executar uma sincronização final. Ao trabalhar com sistemas de arquivos regulares, é fácil garantir que o banco de dados migrado seja 100% sincronizado com o original, pois os logs de refazer originais são copiados e reproduzidos. Não há uma boa maneira de fazer isso com ASM. Apenas os logs de arquivo podem ser facilmente retratados. Para se certificar de que nenhum dado é perdido, o encerramento final do banco de dados original deve ser realizado com cuidado.

1. Primeiro, o banco de dados deve ser silenciado, garantindo que nenhuma alteração esteja sendo feita. Essa quiescência pode incluir a desativação de operações agendadas, o desligamento de ouvintes e/ou o desligamento de aplicativos.
2. Depois que essa etapa é tomada, a maioria dos DBAs cria uma tabela fictícia para servir como um marcador do desligamento.
3. Forçar um arquivo de log para garantir que a criação da tabela fictícia seja gravada nos logs do arquivo. Para fazer isso, execute os seguintes comandos:

```
SQL> create table cutovercheck as select * from dba_users;
Table created.
SQL> alter system archive log current;
System altered.
SQL> shutdown immediate;
Database closed.
Database dismounted.
ORACLE instance shut down.
```

4. Para copiar o último dos registos de arquivo, execute os seguintes comandos. O banco de dados deve estar disponível, mas não aberto.

```
SQL> startup mount;
ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             331353200 bytes
Database Buffers          465567744 bytes
Redo Buffers                5455872 bytes
Database mounted.
```

5. Para copiar os logs de arquivo, execute os seguintes comandos:

```
RMAN> configure channel device type disk format
'/rman/pancake/logship/%h_%e_%a.dbf';
2> backup as copy archivelog from time 'sysdate-2';
3>
4>
using target database control file instead of recovery catalog
old RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters:
CONFIGURE CHANNEL DEVICE TYPE DISK FORMAT
'/rman/pancake/logship/%h_%e_%a.dbf';
new RMAN configuration parameters are successfully stored
Starting backup at 24-MAY-16
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=8 device type=DISK
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=54 RECID=123 STAMP=912659482
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:58:24
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_54_912576125.dbf
continuing other job steps, job failed will not be re-run
...
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=45 RECID=152 STAMP=912659514
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:58:58
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_45_912576125.dbf
continuing other job steps, job failed will not be re-run
channel ORA_DISK_1: starting archived log copy
input archived log thread=1 sequence=47 RECID=153 STAMP=912659515
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03009: failure of backup command on ORA_DISK_1 channel at
05/24/2016 04:59:00
ORA-19635: input and output file names are identical:
/rman/pancake/logship/1_47_912576125.dbf
```

6. Por fim, reproduza os registos de arquivo restantes no novo servidor.

```
[root@jfsc2 pancake]# ./replaylogs.pl PANCAKE
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Tue May 24 05:00:53 2016
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> ORA-00279: change 563137 generated at 05/24/2016 04:36:20 needed
for thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_65_912576125.dbf
ORA-00280: change 563137 for thread 1 is in sequence #65
Specify log: {<RET>=suggested | filename | AUTO | CANCEL}
ORA-00279: change 563629 generated at 05/24/2016 04:55:20 needed for
thread 1
ORA-00289: suggestion : /rman/pancake/logship/1_66_912576125.dbf
ORA-00280: change 563629 for thread 1 is in sequence #66
ORA-00278: log file '/rman/pancake/logship/1_65_912576125.dbf' no longer
needed
for this recovery
ORA-00308: cannot open archived log
'/rman/pancake/logship/1_66_912576125.dbf'
ORA-27037: unable to obtain file status
Linux-x86_64 Error: 2: No such file or directory
Additional information: 3
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
```

7. Nesse estágio, replique todos os dados. O banco de dados está pronto para ser convertido de um banco de dados de reserva para um banco de dados operacional ativo e, em seguida, aberto.

```
SQL> alter database activate standby database;
Database altered.
SQL> alter database open;
Database altered.
```

8. Confirmar a presença da tabela fictícia e, em seguida, soltá-la.

```
SQL> desc cutovercheck
 Name                                      Null?    Type
 ----------------------------------------- --------
---------------------------
 USERNAME                                  NOT NULL VARCHAR2(128)
 USER_ID                                   NOT NULL NUMBER
 PASSWORD                                           VARCHAR2(4000)
 ACCOUNT_STATUS                            NOT NULL VARCHAR2(32)
 LOCK_DATE                                          DATE
 EXPIRY_DATE                                        DATE
 DEFAULT_TABLESPACE                        NOT NULL VARCHAR2(30)
 TEMPORARY_TABLESPACE                      NOT NULL VARCHAR2(30)
 CREATED                                   NOT NULL DATE
 PROFILE                                   NOT NULL VARCHAR2(128)
 INITIAL_RSRC_CONSUMER_GROUP                        VARCHAR2(128)
 EXTERNAL_NAME                                      VARCHAR2(4000)
 PASSWORD_VERSIONS                                  VARCHAR2(12)
 EDITIONS_ENABLED                                   VARCHAR2(1)
 AUTHENTICATION_TYPE                                VARCHAR2(8)
 PROXY_ONLY_CONNECT                                 VARCHAR2(1)
 COMMON                                             VARCHAR2(3)
 LAST_LOGIN                                         TIMESTAMP(9) WITH
TIME ZONE
 ORACLE_MAINTAINED                                  VARCHAR2(1)
SQL> drop table cutovercheck;
Table dropped.
```

REDO log migração sem interrupções

Há momentos em que um banco de dados é organizado corretamente em geral, com exceção dos logs de refazer. Isso pode acontecer por muitos motivos, sendo que o mais comum está relacionado a instantâneos. Produtos como SnapManager para Oracle, SnapCenter e a estrutura de gerenciamento de storage NetApp Snap Creator permitem a recuperação quase instantânea de um banco de dados, mas somente se você reverter o estado dos volumes de arquivos de dados. Se os logs de refazer compartilham espaço com os arquivos de dados, a reversão não pode ser executada com segurança porque isso resultaria na destruição dos logs de refazer, provavelmente significando perda de dados. Portanto, os logs de refazer devem ser relocados.

Esse procedimento é simples e pode ser realizado sem interrupções.

Configuração atual do registo de reprocessamento

1. Identifique o número de grupos de registo de refazer e os respetivos números de grupo.

```
SQL> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
--------------------------------------------------------------------------------
--------
1 /redo0/NTAP/redo01a.log
1 /redo1/NTAP/redo01b.log
2 /redo0/NTAP/redo02a.log
2 /redo1/NTAP/redo02b.log
3 /redo0/NTAP/redo03a.log
3 /redo1/NTAP/redo03b.log
rows selected.
```

2. Introduza o tamanho dos registos de refazer.

```
SQL> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
--------------------------------------------------------------------------------
--------
1 524288000
2 524288000
3 524288000
```

Crie novos logs

1. Para cada log de refazer, crie um novo grupo com um tamanho e número correspondentes de membros.

```
SQL> alter database add logfile ('/newredo0/redo01a.log',
'/newredo1/redo01b.log') size 500M;
Database altered.
SQL> alter database add logfile ('/newredo0/redo02a.log',
'/newredo1/redo02b.log') size 500M;
Database altered.
SQL> alter database add logfile ('/newredo0/redo03a.log',
'/newredo1/redo03b.log') size 500M;
Database altered.
SQL>
```

2. Verifique a nova configuração.

```
SQL> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
--------------------------------------------------------------------------------
--------
1 /redo0/NTAP/redo01a.log
1 /redo1/NTAP/redo01b.log
2 /redo0/NTAP/redo02a.log
2 /redo1/NTAP/redo02b.log
3 /redo0/NTAP/redo03a.log
3 /redo1/NTAP/redo03b.log
4 /newredo0/redo01a.log
4 /newredo1/redo01b.log
5 /newredo0/redo02a.log
5 /newredo1/redo02b.log
6 /newredo0/redo03a.log
6 /newredo1/redo03b.log
12 rows selected.
```

Soltar registos antigos

1. Solte os logs antigos (grupos 1, 2 e 3).

```
SQL> alter database drop logfile group 1;
Database altered.
SQL> alter database drop logfile group 2;
Database altered.
SQL> alter database drop logfile group 3;
Database altered.
```

2. Se encontrar um erro que o impeça de largar um registo ativo, force um interrutor para o registo seguinte para libertar o bloqueio e force um ponto de verificação global. Veja o exemplo a seguir deste processo. A tentativa de soltar o grupo de arquivos de log 2, que estava localizado no local antigo, foi negada porque ainda havia dados ativos neste arquivo de log.

```
SQL> alter database drop logfile group 2;
alter database drop logfile group 2
*
ERROR at line 1:
ORA-01623: log 2 is current log for instance NTAP (thread 1) - cannot
drop
ORA-00312: online log 2 thread 1: '/redo0/NTAP/redo02a.log'
ORA-00312: online log 2 thread 1: '/redo1/NTAP/redo02b.log'
```

3. Um arquivo de log seguido por um ponto de verificação permite que você solte o arquivo de log.

```
SQL> alter system archive log current;
System altered.
SQL> alter system checkpoint;
System altered.
SQL> alter database drop logfile group 2;
Database altered.
```

4. Em seguida, elimine os registos do sistema de ficheiros. Você deve realizar este processo com extremo cuidado.

Cópia de dados do host

Assim como na migração no nível do banco de dados, a migração na camada de host fornece uma abordagem independente de fornecedor de storage.

Em outras palavras, em algum momento "basta copiar os arquivos" é a melhor opção.

Embora essa abordagem de baixa tecnologia possa parecer muito básica, ela oferece benefícios significativos porque nenhum software especial é necessário e os dados originais permanecem intactos com segurança durante o processo. A limitação principal é o fato de que uma migração de dados de cópia de arquivo é um processo disruptivo, porque o banco de dados deve ser desligado antes que a operação de cópia comece. Não há uma boa maneira de sincronizar as alterações dentro de um arquivo, então os arquivos devem ser completamente quiesced antes que a cópia comece.

Se o desligamento exigido por uma operação de cópia não for desejável, a próxima melhor opção baseada em host é aproveitar um gerenciador de volume lógico (LVM). Existem muitas opções de LVM, incluindo Oracle ASM, todas com capacidades semelhantes, mas também com algumas limitações que devem ser levadas em conta. Na maioria dos casos, a migração pode ser realizada sem tempo de inatividade e interrupção.

Filesystem para cópia do sistema de arquivos

A utilidade de uma operação de cópia simples não deve ser subestimada. Essa operação requer tempo de inatividade durante o processo de cópia, mas é um processo altamente confiável e não requer experiência especial com sistemas operacionais, bancos de dados ou sistemas de storage. Além disso, é muito seguro porque não afeta os dados originais. Normalmente, um administrador de sistema altera os sistemas de arquivos de origem para serem montados como somente leitura e, em seguida, reinicia um servidor para garantir que nada pode danificar os dados atuais. O processo de cópia pode ser programado para garantir que ele seja executado o mais rápido possível sem risco de erro do usuário. Como o tipo de e/S é uma transferência sequencial simples de dados, ele é altamente eficiente em largura de banda.

O exemplo a seguir demonstra uma opção para uma migração segura e rápida.

Ambiente

O ambiente a ser migrado é o seguinte:

- Sistemas de arquivos atuais

```
ontap-nfs1:/host1_oradata       52428800  16196928  36231872  31%
/oradata
ontap-nfs1:/host1_logs          49807360    548032  49259328   2% /logs
```

- Novos sistemas de arquivos

```
ontap-nfs1:/host1_logs_new      49807360       128  49807232   1%
/new/logs
ontap-nfs1:/host1_oradata_new   49807360       128  49807232   1%
/new/oradata
```

Visão geral

O banco de dados pode ser migrado por um DBA simplesmente desligando o banco de dados e copiando os arquivos, mas o processo é facilmente programado se muitos bancos de dados precisam ser migrados ou minimizar o tempo de inatividade é fundamental. O uso de scripts também reduz a chance de erro do usuário.

Os scripts de exemplo mostrados automatizam as seguintes operações:

- Encerrar o banco de dados
- Convertendo os sistemas de arquivos existentes em um estado somente leitura
- Copiar todos os dados da origem para os sistemas de arquivos de destino, o que preserva todas as permissões de arquivo
- Desmontar os sistemas de ficheiros antigos e novos
- Remontar os novos sistemas de arquivos nos mesmos caminhos que os sistemas de arquivos anteriores

Procedimento

1. Encerre o banco de dados.

```
[root@host1 current]# ./dbshut.pl NTAP
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 15:58:48 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
```

2. Converta os sistemas de arquivos para somente leitura. Isso pode ser feito mais rapidamente usando um script, como mostrado em "Converter sistema de arquivos para somente leitura".

```
[root@host1 current]# ./mk.fs.readonly.pl /oradata
/oradata unmounted
/oradata mounted read-only
[root@host1 current]# ./mk.fs.readonly.pl /logs
/logs unmounted
/logs mounted read-only
```

3. Confirme se os sistemas de arquivos agora são somente leitura.

```
ontap-nfs1:/host1_oradata on /oradata type nfs
(ro,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
ontap-nfs1:/host1_logs on /logs type nfs
(ro,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
```

4. Sincronize o conteúdo do sistema de arquivos com o `rsync` comando.

```
[root@host1 current]# rsync -rlpogt --stats --progress
--exclude=.snapshot /oradata/ /new/oradata/
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
```

```
 10737426432 100%  153.50MB/s    0:01:06 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%   12.09MB/s    0:00:01 (xfer#2, to-check=9/13)
...
NTAP/undotbs02.dbf
  1073750016 100%  131.60MB/s    0:00:07 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    3.95MB/s    0:00:01 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  162204017.96 bytes/sec
total size is 18570092218  speedup is 1.00
[root@host1 current]# rsync -rlpogt --stats --progress
--exclude=.snapshot /logs/ /new/logs/
sending incremental file list
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%   95.98MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   49.45MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   44.68MB/s    0:00:01 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   68.03MB/s    0:00:00 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
```

```
sent 527098156 bytes  received 278 bytes  95836078.91 bytes/sec
total size is 527032832  speedup is 1.00
```

5. Desmonte os sistemas de arquivos antigos e reposicione os dados copiados. Isso pode ser feito mais rapidamente usando um script, como mostrado em "Substitua o sistema de arquivos".

```
[root@host1 current]# ./swap.fs.pl /logs,/new/logs
/new/logs unmounted
/logs unmounted
Updated /logs mounted
[root@host1 current]# ./swap.fs.pl /oradata,/new/oradata
/new/oradata unmounted
/oradata unmounted
Updated /oradata mounted
```

6. Confirme se os novos sistemas de ficheiros estão na posição correta.

```
ontap-nfs1:/host1_logs_new on /logs type nfs
(rw,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
ontap-nfs1:/host1_oradata_new on /oradata type nfs
(rw,bg,vers=3,rsize=65536,wsize=65536,addr=172.20.101.10)
```

7. Inicie o banco de dados.

```
[root@host1 current]# ./dbstart.pl NTAP
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 16:10:07 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
```

Redução totalmente automatizada

Este script de exemplo aceita argumentos do SID do banco de dados seguido por pares de sistemas de arquivos delimitados por comum. Para o exemplo mostrado acima, o comando é emitido da seguinte forma:

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
```

Quando executado, o script de exemplo tenta executar a seguinte sequência. Ele termina se encontrar um erro em qualquer etapa:

1. Encerre o banco de dados.
2. Converta os sistemas de arquivos atuais para o status somente leitura.
3. Use cada par delimitado por vírgulas de argumentos do sistema de arquivos e sincronize o primeiro sistema de arquivos para o segundo.
4. Desmonte os sistemas de ficheiros anteriores.
5. Atualize o `/etc/fstab` ficheiro da seguinte forma:
   a. Crie uma cópia de segurança em `/etc/fstab.bak`.
   b. Comente as entradas anteriores para os sistemas de ficheiros anteriores e novos.
   c. Crie uma nova entrada para o novo sistema de arquivos que usa o ponto de montagem antigo.
6. Monte os sistemas de ficheiros.
7. Inicie o banco de dados.

O texto a seguir fornece um exemplo de execução para este script:

```
[root@host1 current]# ./migrate.oracle.fs.pl NTAP /logs,/new/logs
/oradata,/new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:05:50 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
SQL> Database closed.
Database dismounted.
ORACLE instance shut down.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP shut down
sending incremental file list
```

```
./
NTAP/
NTAP/1_22_897068759.dbf
    45523968 100%  185.40MB/s    0:00:00 (xfer#1, to-check=15/18)
NTAP/1_23_897068759.dbf
    40601088 100%   81.34MB/s    0:00:00 (xfer#2, to-check=14/18)
...
NTAP/redo/redo02.log
    52429312 100%   70.42MB/s    0:00:00 (xfer#12, to-check=1/18)
NTAP/redo/redo03.log
    52429312 100%   47.08MB/s    0:00:01 (xfer#13, to-check=0/18)
Number of files: 18
Number of files transferred: 13
Total file size: 527032832 bytes
Total transferred file size: 527032832 bytes
Literal data: 527032832 bytes
Matched data: 0 bytes
File list size: 413
File list generation time: 0.001 seconds
File list transfer time: 0.000 seconds
Total bytes sent: 527098156
Total bytes received: 278
sent 527098156 bytes  received 278 bytes  150599552.57 bytes/sec
total size is 527032832  speedup is 1.00
Succesfully replicated filesystem /logs to /new/logs
sending incremental file list
./
NTAP/
NTAP/IOPS.dbf
 10737426432 100%  176.55MB/s    0:00:58 (xfer#1, to-check=10/13)
NTAP/iops.dbf.zip
    22823573 100%    9.48MB/s    0:00:02 (xfer#2, to-check=9/13)
... NTAP/undotbs01.dbf
   309338112 100%   70.76MB/s    0:00:04 (xfer#9, to-check=2/13)
NTAP/undotbs02.dbf
  1073750016 100%  187.65MB/s    0:00:05 (xfer#10, to-check=1/13)
NTAP/users01.dbf
     5251072 100%    5.09MB/s    0:00:00 (xfer#11, to-check=0/13)
Number of files: 13
Number of files transferred: 11
Total file size: 18570092218 bytes
Total transferred file size: 18570092218 bytes
Literal data: 18570092218 bytes
Matched data: 0 bytes
File list size: 277
File list generation time: 0.001 seconds
```

```
File list transfer time: 0.000 seconds
Total bytes sent: 18572359828
Total bytes received: 228
sent 18572359828 bytes  received 228 bytes  177725933.55 bytes/sec
total size is 18570092218  speedup is 1.00
Succesfully replicated filesystem /oradata to /new/oradata
swap 0 /logs /new/logs
/new/logs unmounted
/logs unmounted
Mounted updated /logs
Swapped filesystem /logs for /new/logs
swap 1 /oradata /new/oradata
/new/oradata unmounted
/oradata unmounted
Mounted updated /oradata
Swapped filesystem /oradata for /new/oradata
ORACLE_SID = [oracle] ? The Oracle base has been set to /orabin
SQL*Plus: Release 12.1.0.2.0 Production on Thu Dec 3 17:08:59 2015
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> ORACLE instance started.
Total System Global Area  805306368 bytes
Fixed Size                  2929552 bytes
Variable Size             390073456 bytes
Database Buffers          406847488 bytes
Redo Buffers                5455872 bytes
Database mounted.
Database opened.
SQL> Disconnected from Oracle Database 12c Enterprise Edition Release
12.1.0.2.0 - 64bit Production
With the Partitioning, OLAP, Advanced Analytics and Real Application
Testing options
NTAP started
[root@host1 current]#
```

Migração Oracle ASM spfile e passwd

Uma dificuldade em concluir a migração envolvendo ASM é o arquivo spfile específico ASM e o arquivo de senha. Por padrão, esses arquivos de metadados críticos são criados no primeiro grupo de discos ASM definido. Se um determinado grupo de discos ASM tiver de ser evacuado e removido, o ficheiro spfile e password que regem essa instância ASM deve ser realocado.

Outro caso de uso no qual esses arquivos podem precisar ser relocados é durante a implantação de software de gerenciamento de banco de dados, como o SnapManager para Oracle ou o plug-in SnapCenter Oracle. Um dos recursos desses produtos é restaurar rapidamente um banco de dados revertendo o estado dos LUNs ASM que hospedam os arquivos de dados. Fazer isso requer que o grupo de discos ASM fique offline antes de executar uma restauração. Este não é um problema, desde que os arquivos de dados de um determinado banco de dados sejam isolados em um grupo de discos ASM dedicado.

Quando esse grupo de discos também contém o arquivo ASM spfile/passwd, a única maneira que o grupo de discos pode ser colocado offline é desligar toda a instância ASM. Este é um processo disruptivo, o que significa que o arquivo spfile/passwd precisaria ser relocado.

Ambiente

1. Base de dados SID: TOAST
2. Ficheiros de dados atuais ligados `+DATA`
3. Ficheiros de registo e ficheiros de controlo atuais ligados `+LOGS`
4. Novos grupos de discos ASM estabelecidos como `+NEWDATA` e. `+NEWLOGS`

Localizações de ficheiros ASM spfile/passwd

A realocação desses arquivos pode ser feita sem interrupções. No entanto, por motivos de segurança, a NetApp recomenda desligar o ambiente de banco de dados para que você possa ter certeza de que os arquivos foram realocados e a configuração foi atualizada corretamente. Este procedimento deve ser repetido se várias instâncias ASM estiverem presentes em um servidor.

Identificar instâncias ASM

Identifique as instâncias ASM com base nos dados gravados no `oratab` arquivo. As instâncias ASM são denotadas por um símbolo.

```
-bash-4.1$ cat /etc/oratab | grep '^+'
+ASM:/orabin/grid:N             # line added by Agent
```

Há uma instância ASM chamada ASM neste servidor.

Certifique-se de que todos os bancos de dados estão desligados

O único processo smon visível deve ser o Smon para a instância ASM em uso. A presença de outro processo Smon indica que um banco de dados ainda está em execução.

```
-bash-4.1$ ps -ef | grep smon
oracle     857     1  0 18:26 ?        00:00:00 asm_smon_+ASM
```

O único processo smon é a própria instância ASM. Isso significa que nenhum outro banco de dados está sendo executado e é seguro prosseguir sem o risco de interromper as operações do banco de dados.

Localize arquivos

Identifique a localização atual do arquivo ASM spfile e senha usando os `spget` comandos e. `pwget`

```
bash-4.1$ asmcmd
ASMCMD> spget
+DATA/spfile.ora
```

```
ASMCMD> pwget --asm
+DATA/orapwasm
```

Os arquivos estão localizados na base do `+DATA` grupo de discos.

Copiar ficheiros

Copie os ficheiros para o novo grupo de discos ASM com os `spcopy` comandos e. `pwcopy` Se o novo grupo de discos tiver sido criado recentemente e estiver vazio, poderá ser necessário montar primeiro.

```
ASMCMD> mount NEWDATA
```

```
ASMCMD> spcopy +DATA/spfile.ora +NEWDATA/spfile.ora
copying +DATA/spfile.ora -> +NEWDATA/spfilea.ora
```

```
ASMCMD> pwcopy +DATA/orapwasm +NEWDATA/orapwasm
copying +DATA/orapwasm -> +NEWDATA/orapwasm
```

Os ficheiros foram agora copiados de `+DATA` para `+NEWDATA`.

Atualizar instância ASM

A instância ASM agora deve ser atualizada para refletir a alteração no local. Os `spset` comandos e `pwset` atualizam os metadados ASM necessários para iniciar o grupo de discos ASM.

```
ASMCMD> spset +NEWDATA/spfile.ora
ASMCMD> pwset --asm +NEWDATA/orapwasm
```

Ative ASM usando arquivos atualizados

Neste ponto, a instância ASM ainda usa os locais anteriores desses arquivos. A instância deve ser reiniciada para forçar uma releitura dos arquivos de seus novos locais e liberar bloqueios nos arquivos anteriores.

```
-bash-4.1$ sqlplus / as sysasm
SQL> shutdown immediate;
ASM diskgroups volume disabled
ASM diskgroups dismounted
ASM instance shutdown
```

```
SQL> startup
ASM instance started
Total System Global Area 1140850688 bytes
Fixed Size                  2933400 bytes
Variable Size            1112751464 bytes
ASM Cache                  25165824 bytes
ORA-15032: not all alterations performed
ORA-15017: diskgroup "NEWDATA" cannot be mounted
ORA-15013: diskgroup "NEWDATA" is already mounted
```

Remova arquivos spfile e senhas antigos

Se o procedimento tiver sido executado com êxito, os ficheiros anteriores já não estão bloqueados e podem ser removidos.

```
-bash-4.1$ asmcmd
ASMCMD> rm +DATA/spfile.ora
ASMCMD> rm +DATA/orapwasm
```

Cópia Oracle ASM para ASM

O Oracle ASM é essencialmente um gerenciador de volumes e um sistema de arquivos combinados leves. Como o sistema de arquivos não é facilmente visível, o RMAN deve ser usado para executar operações de cópia. Embora um processo de migração baseado em cópia seja seguro e simples, isso resulta em algumas interrupções. A interrupção pode ser minimizada, mas não totalmente eliminada.

Se você quiser migração sem interrupções de um banco de dados baseado em ASM, a melhor opção é utilizar a funcionalidade do ASM para rebalancear as extensões ASM para novos LUNs e deixar cair os LUNs antigos. Isso geralmente é seguro e sem interrupções para operações, mas não oferece caminho de back-out. Se forem encontrados problemas funcionais ou de desempenho, a única opção é migrar os dados de volta para a origem.

Esse risco pode ser evitado copiando o banco de dados para o novo local em vez de mover dados, para que os dados originais fiquem intactos. O banco de dados pode ser totalmente testado em seu novo local antes de entrar em funcionamento, e o banco de dados original está disponível como uma opção de retorno se problemas forem encontrados.

Este procedimento é uma das muitas opções envolvendo RMAN. Ele foi projetado para permitir um processo de duas etapas no qual o backup inicial é criado e, em seguida, sincronizado mais tarde através da repetição de log. Esse processo é desejável para minimizar o tempo de inatividade, pois permite que o banco de dados permaneça operacional e forneça dados durante a cópia inicial da linha de base.

Copiar base de dados

O Oracle RMAN cria uma cópia de nível 0 (completa) do banco de dados de origem atualmente localizado no grupo de discos ASM `+DATA` para o novo local no `+NEWDATA`.

```
-bash-4.1$ rman target /
Recovery Manager: Release 12.1.0.2.0 - Production on Sun Dec 6 17:40:03
2015
Copyright (c) 1982, 2014, Oracle and/or its affiliates.  All rights
reserved.
connected to target database: TOAST (DBID=2084313411)
RMAN> backup as copy incremental level 0 database format '+NEWDATA' tag
'ONTAP_MIGRATION';
Starting backup at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=302 device type=DISK
channel ORA_DISK_1: starting datafile copy
input datafile file number=00001
name=+DATA/TOAST/DATAFILE/system.262.897683141
...
input datafile file number=00004
name=+DATA/TOAST/DATAFILE/users.264.897683151
output file name=+NEWDATA/TOAST/DATAFILE/users.258.897759623
tag=ONTAP_MIGRATION RECID=5 STAMP=897759622
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 0 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWDATA/TOAST/BACKUPSET/2015_12_06/nnsnn0_ontap_migration_0.262.89
7759623 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

Forçar o interrutor de registo de arquivo

Você deve forçar um switch de log de arquivamento para garantir que os logs de arquivamento contenham todos os dados necessários para tornar a cópia totalmente consistente. Sem este comando, os dados de chave ainda podem estar presentes nos logs de refazer.

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Desligue o banco de dados de origem

A interrupção começa nesta etapa porque o banco de dados é desligado e colocado em um modo de acesso limitado, somente leitura. Para encerrar o banco de dados de origem, execute os seguintes comandos:

```
RMAN> shutdown immediate;
using target database control file instead of recovery catalog
database closed
database dismounted
Oracle instance shut down
RMAN> startup mount;
connected to target database (not started)
Oracle instance started
database mounted
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                390073456 bytes
Database Buffers             406847488 bytes
Redo Buffers                   5455872 bytes
```

Backup do Controlfile

Você deve fazer backup do controlfile caso precise abortar a migração e reverter para o local de armazenamento original. Uma cópia do ficheiro de controlo de cópia de segurança não é 100% necessária, mas facilita o processo de reposição das localizações dos ficheiros de base de dados para a localização original.

```
RMAN> backup as copy current controlfile format '/tmp/TOAST.ctrl';
Starting backup at 06-DEC-15
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=358 device type=DISK
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/tmp/TOAST.ctrl tag=TAG20151206T174753 RECID=6
STAMP=897760073
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

Atualizações de parâmetros

O spfile atual contém referências aos controlfiles em seus locais atuais dentro do grupo de discos ASM antigo. Ele deve ser editado, o que é facilmente feito editando uma versão intermediária pfile.

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Atualize o pfile

Atualize quaisquer parâmetros referentes a grupos de discos ASM antigos para refletir os novos nomes de grupos de discos ASM. Em seguida, salve o arquivo pfile atualizado. Certifique-se de que os `db_create`

parâmetros estão presentes.

No exemplo abaixo, as referências a `+DATA` que foram alteradas `+NEWDATA` são realçadas em amarelo. Dois parâmetros-chave são os `db_create` parâmetros que criam quaisquer novos arquivos no local correto.

```
*.compatible='12.1.0.2.0'
*.control_files='+NEWLOGS/TOAST/CONTROLFILE/current.258.897683139'
*.db_block_size=8192
*. db_create_file_dest='+NEWDATA'
*. db_create_online_log_dest_1='+NEWLOGS'
*.db_domain=''
*.db_name='TOAST'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=TOASTXDB)'
*.log_archive_dest_1='LOCATION=+NEWLOGS'
*.log_archive_format='%t_%s_%r.dbf'
```

Atualize o arquivo init.ora

A maioria dos bancos de dados baseados em ASM usa um `init.ora` arquivo localizado no `$ORACLE_HOME/dbs` diretório, que é um ponto para o spfile no grupo de discos ASM. Esse arquivo deve ser redirecionado para um local no novo grupo de discos ASM.

```
-bash-4.1$ cd $ORACLE_HOME/dbs
-bash-4.1$ cat initTOAST.ora
SPFILE='+DATA/TOAST/spfileTOAST.ora'
```

Altere este ficheiro da seguinte forma:

```
SPFILE=+NEWLOGS/TOAST/spfileTOAST.ora
```

Recriação do arquivo de parâmetros

O arquivo spfile agora está pronto para ser preenchido pelos dados no arquivo pfile editado.

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Inicie o banco de dados para começar a usar o novo spfile

Inicie o banco de dados para se certificar de que ele agora usa o arquivo spfile recém-criado e que quaisquer outras alterações aos parâmetros do sistema são registradas corretamente.

```
RMAN> startup nomount;
connected to target database (not started)
Oracle instance started
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                373296240 bytes
Database Buffers             423624704 bytes
Redo Buffers                   5455872 bytes
```

Restaure o ficheiro de controlo

O arquivo de controle de backup criado pelo RMAN também pode ser restaurado pelo RMAN diretamente para o local especificado no novo spfile.

```
RMAN> restore controlfile from
'+DATA/TOAST/CONTROLFILE/current.258.897683139';
Starting restore at 06-DEC-15
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=417 device type=DISK
channel ORA_DISK_1: copied control file copy
output file name=+NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
Finished restore at 06-DEC-15
```

Monte o banco de dados e verifique o uso do novo controlfile.

```
RMAN> alter database mount;
using target database control file instead of recovery catalog
Statement processed
```

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
+NEWLOGS/TOAST/CONTROLFILE/cur
                                                 rent.273.897761061
```

Registo de reprodução

O banco de dados usa atualmente os arquivos de dados no local antigo. Antes que a cópia possa ser usada, ela deve ser sincronizada. O tempo passou durante o processo de cópia inicial e as alterações foram registradas principalmente nos logs de arquivo. Essas alterações são replicadas da seguinte forma:

1. Execute uma cópia de segurança incremental RMAN, que contém os registos de arquivo.

```
RMAN> backup incremental level 1 format '+NEWLOGS' for recover of copy
with tag 'ONTAP_MIGRATION' database;
Starting backup at 06-DEC-15
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=62 device type=DISK
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001
name=+DATA/TOAST/DATAFILE/system.262.897683141
input datafile file number=00002
name=+DATA/TOAST/DATAFILE/sysaux.260.897683143
input datafile file number=00003
name=+DATA/TOAST/DATAFILE/undotbs1.257.897683145
input datafile file number=00004
name=+DATA/TOAST/DATAFILE/users.264.897683151
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.
897762693 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current control file in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 06-DEC-15
channel ORA_DISK_1: finished piece 1 at 06-DEC-15
piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/ncsnn1_ontap_migration_0.267.
897762697 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 06-DEC-15
```

2. Repetir o registo.

```
RMAN> recover copy of database with tag 'ONTAP_MIGRATION';
Starting recover at 06-DEC-15
using channel ORA_DISK_1
channel ORA_DISK_1: starting incremental datafile backup set restore
channel ORA_DISK_1: specifying datafile copies to recover
recovering datafile copy file number=00001
name=+NEWDATA/TOAST/DATAFILE/system.259.897759609
recovering datafile copy file number=00002
name=+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615
recovering datafile copy file number=00003
name=+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619
recovering datafile copy file number=00004
name=+NEWDATA/TOAST/DATAFILE/users.258.897759623
channel ORA_DISK_1: reading from backup piece
+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.8977626
93
channel ORA_DISK_1: piece
handle=+NEWLOGS/TOAST/BACKUPSET/2015_12_06/nnndn1_ontap_migration_0.268.
897762693 tag=ONTAP_MIGRATION
channel ORA_DISK_1: restored backup piece 1
channel ORA_DISK_1: restore complete, elapsed time: 00:00:01
Finished recover at 06-DEC-15
```

Ativação

O arquivo de controle que foi restaurado ainda faz referência aos arquivos de dados no local original e também contém as informações de caminho para os arquivos de dados copiados.

1. Para alterar os arquivos de dados ativos, execute o `switch database to copy` comando.

```
RMAN> switch database to copy;
datafile 1 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/system.259.897759609"
datafile 2 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615"
datafile 3 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619"
datafile 4 switched to datafile copy
"+NEWDATA/TOAST/DATAFILE/users.258.897759623"
```

   Os arquivos de dados ativos agora são os arquivos de dados copiados, mas ainda podem haver alterações contidas nos logs de refazer finais.

2. Para reproduzir todos os logs restantes, execute o `recover database` comando. Se a mensagem `media recovery complete` for exibida, o processo foi bem-sucedido.

```
RMAN> recover database;
Starting recover at 06-DEC-15
using channel ORA_DISK_1
starting media recovery
media recovery complete, elapsed time: 00:00:01
Finished recover at 06-DEC-15
```

Este processo só alterou a localização dos ficheiros de dados normais. Os arquivos de dados temporários devem ser renomeados, mas não precisam ser copiados porque são apenas temporários. O banco de dados está inativo no momento, portanto não há dados ativos nos arquivos de dados temporários.

3. Para realocar os arquivos de dados temporários, primeiro identifique sua localização.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
--------------------------------------------------------------------------------
--------
1 +DATA/TOAST/TEMPFILE/temp.263.897683145
```

4. Reposicione arquivos de dados temporários usando um comando RMAN que define o novo nome para cada arquivo de dados. Com o Oracle Managed Files (OMF), o nome completo não é necessário; o grupo de discos ASM é suficiente. Quando o banco de dados é aberto, o OMF vincula ao local apropriado no grupo de discos ASM. Para realocar arquivos, execute os seguintes comandos:

```
run {
set newname for tempfile 1 to '+NEWDATA';
switch tempfile all;
}
```

```
RMAN> run {
2> set newname for tempfile 1 to '+NEWDATA';
3> switch tempfile all;
4> }
executing command: SET NEWNAME
renamed tempfile 1 to +NEWDATA in control file
```

Refazer a migração de log

O processo de migração está quase concluído, mas os logs de refazer ainda estão localizados no grupo de discos ASM original. Os registos de refazer não podem ser transferidos diretamente. Em vez disso, um novo conjunto de logs de refazer é criado e adicionado à configuração, seguido de uma gota dos logs antigos.

1. Identifique o número de grupos de registo de refazer e os respetivos números de grupo.

```
RMAN> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
------------------------------------------------------------------------------
--------
1 +DATA/TOAST/ONLINELOG/group_1.261.897683139
2 +DATA/TOAST/ONLINELOG/group_2.259.897683139
3 +DATA/TOAST/ONLINELOG/group_3.256.897683139
```

2. Introduza o tamanho dos registos de refazer.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
------------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. Para cada log refazer, crie um novo grupo com uma configuração correspondente. Se você não estiver usando OMF, você deve especificar o caminho completo. Este também é um exemplo que usa os `db_create_online_log` parâmetros. Como foi mostrado anteriormente, este parâmetro foi definido para -NEWLOGS. Esta configuração permite que você use os seguintes comandos para criar novos logs on-line sem a necessidade de especificar um local de arquivo ou mesmo um grupo de discos ASM específico.

```
RMAN> alter database add logfile size 52428800;
Statement processed
RMAN> alter database add logfile size 52428800;
Statement processed
RMAN> alter database add logfile size 52428800;
Statement processed
```

4. Abra o banco de dados.

```
SQL> alter database open;
Database altered.
```

5. Solte os logs antigos.

```
RMAN> alter database drop logfile group 1;
Statement processed
```

6. Se encontrar um erro que o impeça de largar um registo ativo, force um interrutor para o registo seguinte para libertar o bloqueio e force um ponto de verificação global. Um exemplo é mostrado abaixo. A tentativa de soltar o grupo de arquivos de log 3, que estava localizado no local antigo, foi negada porque ainda havia dados ativos neste arquivo de log. Um arquivo de log após um ponto de verificação permite que você exclua o arquivo de log.

```
RMAN> alter database drop logfile group 3;
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03002: failure of sql statement command at 12/08/2015 20:23:51
ORA-01623: log 3 is current log for instance TOAST (thread 4) - cannot
drop
ORA-00312: online log 3 thread 1:
'+LOGS/TOAST/ONLINELOG/group_3.259.897563549'
RMAN> alter system switch logfile;
Statement processed
RMAN> alter system checkpoint;
Statement processed
RMAN> alter database drop logfile group 3;
Statement processed
```

7. Revise o ambiente para garantir que todos os parâmetros baseados em localização sejam atualizados.

```
SQL> select name from v$datafile;
SQL> select member from v$logfile;
SQL> select name from v$tempfile;
SQL> show parameter spfile;
SQL> select name, value from v$parameter where value is not null;
```

8. O script a seguir demonstra como simplificar esse processo:

```
[root@host1 current]# ./checkdbdata.pl TOAST
TOAST datafiles:
+NEWDATA/TOAST/DATAFILE/system.259.897759609
+NEWDATA/TOAST/DATAFILE/sysaux.263.897759615
+NEWDATA/TOAST/DATAFILE/undotbs1.264.897759619
+NEWDATA/TOAST/DATAFILE/users.258.897759623
TOAST redo logs:
+NEWLOGS/TOAST/ONLINELOG/group_4.266.897763123
+NEWLOGS/TOAST/ONLINELOG/group_5.265.897763125
+NEWLOGS/TOAST/ONLINELOG/group_6.264.897763125
TOAST temp datafiles:
+NEWDATA/TOAST/TEMPFILE/temp.260.897763165
TOAST spfile
spfile                                string
+NEWDATA/spfiletoast.ora
TOAST key parameters
control_files +NEWLOGS/TOAST/CONTROLFILE/current.273.897761061
log_archive_dest_1 LOCATION=+NEWLOGS
db_create_file_dest +NEWDATA
db_create_online_log_dest_1 +NEWLOGS
```

9. Se os grupos de discos ASM foram completamente evacuados, eles agora podem ser desmontados com `asmcmd`. No entanto, em muitos casos, os arquivos pertencentes a outros bancos de dados ou o arquivo ASM spfile/passwd ainda podem estar presentes.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Oracle ASM para cópia do sistema de arquivos

O procedimento de cópia do Oracle ASM para sistema de arquivos é muito semelhante ao procedimento de cópia ASM para ASM, com benefícios e restrições semelhantes. A principal diferença é a sintaxe dos vários comandos e parâmetros de configuração ao usar um sistema de arquivos visível em vez de um grupo de discos ASM.

Copiar base de dados

O Oracle RMAN é usado para criar uma cópia de nível 0 (completa) do banco de dados de origem atualmente localizado no grupo de discos ASM `+DATA` para o novo local no `/oradata`.

```
RMAN> backup as copy incremental level 0 database format
'/oradata/TOAST/%U' tag 'ONTAP_MIGRATION';
Starting backup at 13-MAY-16
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=377 device type=DISK
channel ORA_DISK_1: starting datafile copy
input datafile file number=00001 name=+ASM0/TOAST/system01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-
1_01r5fhjg tag=ONTAP_MIGRATION RECID=1 STAMP=911722099
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
input datafile file number=00002 name=+ASM0/TOAST/sysaux01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-
2_02r5fhjo tag=ONTAP_MIGRATION RECID=2 STAMP=911722106
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
input datafile file number=00003 name=+ASM0/TOAST/undotbs101.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt tag=ONTAP_MIGRATION RECID=3 STAMP=911722113
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:07
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/oradata/TOAST/cf_D-TOAST_id-2098173325_04r5fhk5
tag=ONTAP_MIGRATION RECID=4 STAMP=911722118
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting datafile copy
input datafile file number=00004 name=+ASM0/TOAST/users01.dbf
output file name=/oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-
4_05r5fhk6 tag=ONTAP_MIGRATION RECID=5 STAMP=911722118
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
channel ORA_DISK_1: starting incremental level 0 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
including current SPFILE in backup set
channel ORA_DISK_1: starting piece 1 at 13-MAY-16
channel ORA_DISK_1: finished piece 1 at 13-MAY-16
piece handle=/oradata/TOAST/06r5fhk7_1_1 tag=ONTAP_MIGRATION comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 13-MAY-16
```

Forçar o interrutor de registo de arquivo

É necessário forçar o comutador de registo de arquivo para garantir que os registos de arquivo contêm todos os dados necessários para tornar a cópia totalmente consistente. Sem este comando, os dados de chave ainda podem estar presentes nos logs de refazer. Para forçar um switch de log de arquivo, execute o seguinte comando:

```
RMAN> sql 'alter system archive log current';
sql statement: alter system archive log current
```

Desligue o banco de dados de origem

A interrupção começa nesta etapa porque o banco de dados é desligado e colocado em um modo somente leitura de acesso limitado. Para encerrar o banco de dados de origem, execute os seguintes comandos:

```
RMAN> shutdown immediate;
using target database control file instead of recovery catalog
database closed
database dismounted
Oracle instance shut down
RMAN> startup mount;
connected to target database (not started)
Oracle instance started
database mounted
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                331353200 bytes
Database Buffers             465567744 bytes
Redo Buffers                   5455872 bytes
```

Backup do Controlfile

Faça backup de arquivos de controle no caso de você precisar abortar a migração e reverter para o local de armazenamento original. Uma cópia do ficheiro de controlo de cópia de segurança não é 100% necessária, mas facilita o processo de reposição das localizações dos ficheiros de base de dados para a localização original.

```
RMAN> backup as copy current controlfile format '/tmp/TOAST.ctrl';
Starting backup at 08-DEC-15
using channel ORA_DISK_1
channel ORA_DISK_1: starting datafile copy
copying current control file
output file name=/tmp/TOAST.ctrl tag=TAG20151208T194540 RECID=30
STAMP=897939940
channel ORA_DISK_1: datafile copy complete, elapsed time: 00:00:01
Finished backup at 08-DEC-15
```

Atualizações de parâmetros

```
RMAN> create pfile='/tmp/pfile' from spfile;
Statement processed
```

Atualize o pfile

Quaisquer parâmetros referentes a grupos de discos ASM antigos devem ser atualizados e, em alguns casos, excluídos quando não forem mais relevantes. Atualize-os para refletir os novos caminhos do sistema de arquivos e salvar o arquivo pfile atualizado. Certifique-se de que o caminho de destino completo está listado. Para atualizar esses parâmetros, execute os seguintes comandos:

```
*.audit_file_dest='/orabin/admin/TOAST/adump'
*.audit_trail='db'
*.compatible='12.1.0.2.0'
*.control_files='/logs/TOAST/arch/control01.ctl','/logs/TOAST/redo/control
02.ctl'
*.db_block_size=8192
*.db_domain=''
*.db_name='TOAST'
*.diagnostic_dest='/orabin'
*.dispatchers='(PROTOCOL=TCP) (SERVICE=TOASTXDB)'
*.log_archive_dest_1='LOCATION=/logs/TOAST/arch'
*.log_archive_format='%t_%s_%r.dbf'
*.open_cursors=300
*.pga_aggregate_target=256m
*.processes=300
*.remote_login_passwordfile='EXCLUSIVE'
*.sga_target=768m
*.undo_tablespace='UNDOTBS1'
```

Desative o arquivo init.ora original

Esse arquivo está localizado no `$ORACLE_HOME/dbs` diretório e geralmente está em um arquivo pfile que serve como um ponteiro para o arquivo spfile no grupo de discos ASM. Para se certificar de que o arquivo spfile original não é mais usado, renomeie-o. Não o exclua, no entanto, porque esse arquivo é necessário se a migração tiver de ser abortada.

```
[oracle@jfsc1 ~]$ cd $ORACLE_HOME/dbs
[oracle@jfsc1 dbs]$ cat initTOAST.ora
SPFILE='+ASM0/TOAST/spfileTOAST.ora'
[oracle@jfsc1 dbs]$ mv initTOAST.ora initTOAST.ora.prev
[oracle@jfsc1 dbs]$
```

Recriação do arquivo de parâmetros

Esta é a etapa final na realocação do arquivo spfile. O arquivo spfile original não é mais usado e o banco de dados está atualmente iniciado (mas não montado) usando o arquivo intermediário. O conteúdo deste arquivo pode ser escrito para o novo local spfile da seguinte forma:

```
RMAN> create spfile from pfile='/tmp/pfile';
Statement processed
```

Inicie o banco de dados para começar a usar o novo spfile

Você deve iniciar o banco de dados para liberar os bloqueios no arquivo intermediário e iniciar o banco de dados usando apenas o novo arquivo spfile. Iniciar o banco de dados também prova que o novo local spfile está correto e seus dados são válidos.

```
RMAN> shutdown immediate;
Oracle instance shut down
RMAN> startup nomount;
connected to target database (not started)
Oracle instance started
Total System Global Area     805306368 bytes
Fixed Size                     2929552 bytes
Variable Size                331353200 bytes
Database Buffers             465567744 bytes
Redo Buffers                   5455872 bytes
```

Restaure o ficheiro de controlo

Um arquivo de controle de backup foi criado no caminho `/tmp/TOAST.ctrl` anteriormente no procedimento. O novo spfile define os locais do controlfile como /`logfs/TOAST/ctrl/ctrlfile1.ctrl` e `/logfs/TOAST/redo/ctrlfile2.ctrl`. No entanto, esses arquivos ainda não existem.

1. Este comando restaura os dados do controlfile para os caminhos definidos no spfile.

   ```
   RMAN> restore controlfile from '/tmp/TOAST.ctrl';
   Starting restore at 13-MAY-16
   using channel ORA_DISK_1
   channel ORA_DISK_1: copied control file copy
   output file name=/logs/TOAST/arch/control01.ctl
   output file name=/logs/TOAST/redo/control02.ctl
   Finished restore at 13-MAY-16
   ```

2. Emita o comando mount para que os controlfiles sejam descobertos corretamente e contenham dados válidos.

```
RMAN> alter database mount;
Statement processed
released channel: ORA_DISK_1
```

Para validar o `control_files` parâmetro, execute o seguinte comando:

```
SQL> show parameter control_files;
NAME                                 TYPE        VALUE
------------------------------------ -----------
------------------------------
control_files                        string
/logs/TOAST/arch/control01.ctl
                                                 ,
/logs/TOAST/redo/control02.c
                                                 tl
```

Registo de reprodução

O banco de dados está usando os arquivos de dados no local antigo. Antes que a cópia possa ser usada, os arquivos de dados devem ser sincronizados. O tempo passou durante o processo de cópia inicial, e as alterações foram registradas principalmente nos Registros de arquivamento. Essas alterações são replicadas nas duas etapas a seguir.

1. Execute uma cópia de segurança incremental RMAN, que contém os registos de arquivo.

```
RMAN>  backup incremental level 1 format '/logs/TOAST/arch/%U' for
recover of copy with tag 'ONTAP_MIGRATION' database;
Starting backup at 13-MAY-16
using target database control file instead of recovery catalog
allocated channel: ORA_DISK_1
channel ORA_DISK_1: SID=124 device type=DISK
channel ORA_DISK_1: starting incremental level 1 datafile backup set
channel ORA_DISK_1: specifying datafile(s) in backup set
input datafile file number=00001 name=+ASM0/TOAST/system01.dbf
input datafile file number=00002 name=+ASM0/TOAST/sysaux01.dbf
input datafile file number=00003 name=+ASM0/TOAST/undotbs101.dbf
input datafile file number=00004 name=+ASM0/TOAST/users01.dbf
channel ORA_DISK_1: starting piece 1 at 13-MAY-16
channel ORA_DISK_1: finished piece 1 at 13-MAY-16
piece handle=/logs/TOAST/arch/09r5fj8i_1_1 tag=ONTAP_MIGRATION
comment=NONE
channel ORA_DISK_1: backup set complete, elapsed time: 00:00:01
Finished backup at 13-MAY-16
RMAN-06497: WARNING: control file is not current, control file
AUTOBACKUP skipped
```

2. Repetir os registos.

```
RMAN> recover copy of database with tag 'ONTAP_MIGRATION';
Starting recover at 13-MAY-16
using channel ORA_DISK_1
channel ORA_DISK_1: starting incremental datafile backup set restore
channel ORA_DISK_1: specifying datafile copies to recover
recovering datafile copy file number=00001 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
recovering datafile copy file number=00002 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
recovering datafile copy file number=00003 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt
recovering datafile copy file number=00004 name=/oradata/TOAST/data_D-
TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
channel ORA_DISK_1: reading from backup piece
/logs/TOAST/arch/09r5fj8i_1_1
channel ORA_DISK_1: piece handle=/logs/TOAST/arch/09r5fj8i_1_1
tag=ONTAP_MIGRATION
channel ORA_DISK_1: restored backup piece 1
channel ORA_DISK_1: restore complete, elapsed time: 00:00:01
Finished recover at 13-MAY-16
RMAN-06497: WARNING: control file is not current, control file
AUTOBACKUP skipped
```

Ativação

O arquivo de controle que foi restaurado ainda faz referência aos arquivos de dados no local original e também contém as informações de caminho para os arquivos de dados copiados.

1. Para alterar os arquivos de dados ativos, execute o `switch database to copy` comando:

```
RMAN> switch database to copy;
datafile 1 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-SYSTEM_FNO-1_01r5fhjg"
datafile 2 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-SYSAUX_FNO-2_02r5fhjo"
datafile 3 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt"
datafile 4 switched to datafile copy "/oradata/TOAST/data_D-TOAST_I-
2098173325_TS-USERS_FNO-4_05r5fhk6"
```

2. Embora os arquivos de dados devam ser totalmente consistentes, uma etapa final é necessária para reproduzir as alterações restantes registradas nos logs de refazer on-line. Use o `recover database` comando para reproduzir essas alterações e tornar a cópia 100% idêntica ao original. No entanto, a cópia ainda não está aberta.

```
RMAN> recover database;
Starting recover at 13-MAY-16
using channel ORA_DISK_1
starting media recovery
archived log for thread 1 with sequence 28 is already on disk as file
+ASM0/TOAST/redo01.log
archived log file name=+ASM0/TOAST/redo01.log thread=1 sequence=28
media recovery complete, elapsed time: 00:00:00
Finished recover at 13-MAY-16
```

Realocar arquivos de dados temporários

1. Identificar a localização dos ficheiros de dados temporários ainda em utilização no grupo de discos original.

```
RMAN> select file#||' '||name from v$tempfile;
FILE#||''||NAME
-------------------------------------------------------------------------
--------
1 +ASM0/TOAST/temp01.dbf
```

2. Para realocar os arquivos de dados, execute os seguintes comandos. Se houver muitos tempfiles, use um editor de texto para criar o comando RMAN e, em seguida, corte e cole-o.

```
RMAN> run {
2> set newname for tempfile 1 to '/oradata/TOAST/temp01.dbf';
3> switch tempfile all;
4> }
executing command: SET NEWNAME
renamed tempfile 1 to /oradata/TOAST/temp01.dbf in control file
```

Refazer a migração de log

O processo de migração está quase concluído, mas os logs de refazer ainda estão localizados no grupo de discos ASM original. Os registos de refazer não podem ser transferidos diretamente. Em vez disso, um novo conjunto de logs de refazer é criado e adicionado à configuração, seguindo-se uma gota dos logs antigos.

1. Identifique o número de grupos de registo de refazer e os respetivos números de grupo.

```
RMAN> select group#||' '||member from v$logfile;
GROUP#||''||MEMBER
------------------------------------------------------------------------
--------
1 +ASM0/TOAST/redo01.log
2 +ASM0/TOAST/redo02.log
3 +ASM0/TOAST/redo03.log
```

2. Introduza o tamanho dos registos de refazer.

```
RMAN> select group#||' '||bytes from v$log;
GROUP#||''||BYTES
------------------------------------------------------------------------
--------
1 52428800
2 52428800
3 52428800
```

3. Para cada log de refazer, crie um novo grupo usando o mesmo tamanho que o grupo de log de refazer atual usando o novo local do sistema de arquivos.

```
RMAN> alter database add logfile '/logs/TOAST/redo/log00.rdo' size
52428800;
Statement processed
RMAN> alter database add logfile '/logs/TOAST/redo/log01.rdo' size
52428800;
Statement processed
RMAN> alter database add logfile '/logs/TOAST/redo/log02.rdo' size
52428800;
Statement processed
```

4. Remova os grupos de arquivos de log antigos que ainda estão localizados no armazenamento anterior.

```
RMAN> alter database drop logfile group 4;
Statement processed
RMAN> alter database drop logfile group 5;
Statement processed
RMAN> alter database drop logfile group 6;
Statement processed
```

5. Se for encontrado um erro que bloqueia a queda de um log ativo, force um switch para o próximo log para liberar o bloqueio e forçar um ponto de verificação global. Um exemplo é mostrado abaixo. A tentativa de soltar o grupo de arquivos de log 3, que estava localizado no local antigo, foi negada porque ainda havia

dados ativos neste arquivo de log. Um arquivo de log seguido por um ponto de verificação permite a exclusão de arquivos de log.

```
RMAN> alter database drop logfile group 4;
RMAN-00571: ===========================================================
RMAN-00569: =============== ERROR MESSAGE STACK FOLLOWS ===============
RMAN-00571: ===========================================================
RMAN-03002: failure of sql statement command at 12/08/2015 20:23:51
ORA-01623: log 4 is current log for instance TOAST (thread 4) - cannot
drop
ORA-00312: online log 4 thread 1:
'+NEWLOGS/TOAST/ONLINELOG/group_4.266.897763123'
RMAN> alter system switch logfile;
Statement processed
RMAN> alter system checkpoint;
Statement processed
RMAN> alter database drop logfile group 4;
Statement processed
```

6. Revise o ambiente para garantir que todos os parâmetros baseados em localização sejam atualizados.

```
SQL> select name from v$datafile;
SQL> select member from v$logfile;
SQL> select name from v$tempfile;
SQL> show parameter spfile;
SQL> select name, value from v$parameter where value is not null;
```

7. O script a seguir demonstra como tornar esse processo mais fácil.

```
[root@jfsc1 current]# ./checkdbdata.pl TOAST
TOAST datafiles:
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-3_03r5fhjt
/oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
TOAST redo logs:
/logs/TOAST/redo/log00.rdo
/logs/TOAST/redo/log01.rdo
/logs/TOAST/redo/log02.rdo
TOAST temp datafiles:
/oradata/TOAST/temp01.dbf
TOAST spfile
spfile                                  string
/orabin/product/12.1.0/dbhome_
                                                      1/dbs/spfileTOAST.ora
TOAST key parameters
control_files /logs/TOAST/arch/control01.ctl,
/logs/TOAST/redo/control02.ctl
log_archive_dest_1 LOCATION=/logs/TOAST/arch
```

8. Se os grupos de discos ASM foram completamente evacuados, eles agora podem ser desmontados com `asmcmd`. Em muitos casos, os arquivos pertencentes a outros bancos de dados ou o arquivo ASM spfile/passwd ainda podem estar presentes.

```
-bash-4.1$ . oraenv
ORACLE_SID = [TOAST] ? +ASM
The Oracle base remains unchanged with value /orabin
-bash-4.1$ asmcmd
ASMCMD> umount DATA
ASMCMD>
```

Procedimento de limpeza do ficheiro de dados

O processo de migração pode resultar em arquivos de dados com sintaxe longa ou críptica, dependendo de como o Oracle RMAN foi usado. No exemplo mostrado aqui, o backup foi realizado com o formato de arquivo `/oradata/TOAST/%U` do . `%U` Indica que o RMAN deve criar um nome exclusivo padrão para cada arquivo de dados. O resultado é semelhante ao que é mostrado no texto a seguir. Os nomes tradicionais para os arquivos de dados são incorporados nos nomes. Isso pode ser limpo usando a abordagem roteirizada mostrada em "Limpeza de migração ASM".

```
[root@jfsc1 current]# ./fixuniquenames.pl TOAST
#sqlplus Commands
shutdown immediate;
startup mount;
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSTEM_FNO-1_01r5fhjg
/oradata/TOAST/system.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-SYSAUX_FNO-2_02r5fhjo
/oradata/TOAST/sysaux.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-UNDOTBS1_FNO-
3_03r5fhjt /oradata/TOAST/undotbs1.dbf
host mv /oradata/TOAST/data_D-TOAST_I-2098173325_TS-USERS_FNO-4_05r5fhk6
/oradata/TOAST/users.dbf
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSTEM_FNO-1_01r5fhjg' to '/oradata/TOAST/system.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
SYSAUX_FNO-2_02r5fhjo' to '/oradata/TOAST/sysaux.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
UNDOTBS1_FNO-3_03r5fhjt' to '/oradata/TOAST/undotbs1.dbf';
alter database rename file '/oradata/TOAST/data_D-TOAST_I-2098173325_TS-
USERS_FNO-4_05r5fhk6' to '/oradata/TOAST/users.dbf';
alter database open;
```

Rebalancear o Oracle ASM

Como discutido anteriormente, um grupo de discos Oracle ASM pode ser migrado de forma transparente para um novo sistema de storage usando o processo de rebalanceamento. Em resumo, o processo de rebalanceamento requer a adição de LUNs de tamanho igual ao grupo de LUNs existente, seguido de uma operação de queda do LUN anterior. O Oracle ASM relocaliza automaticamente os dados subjacentes para o novo storage em um layout ideal e, em seguida, libera os LUNs antigos quando concluído.

O processo de migração usa e/S sequenciais eficientes e geralmente não causa interrupções no desempenho, mas a taxa de migração pode ser controlada quando necessário.

Identifique os dados a serem migrados

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
NEWDATA_0003 1 10240 /dev/mapper/3600a090038303537762b47594c315864 MEMBER
NEWDATA_0002 1 10240 /dev/mapper/3600a090038303537762b47594c315863 MEMBER
NEWDATA_0000 1 10240 /dev/mapper/3600a090038303537762b47594c315861 MEMBER
NEWDATA_0001 1 10240 /dev/mapper/3600a090038303537762b47594c315862 MEMBER
SQL> select group_number||' '||name from v$asm_diskgroup;
1 NEWDATA
```

Criar novos LUNs

Crie novos LUNs do mesmo tamanho e defina a associação de usuário e grupo conforme necessário. Os LUNs devem aparecer como `CANDIDATE` discos.

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
 0 0 /dev/mapper/3600a098038303537762b47594c31586b CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c315869 CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c315858 CANDIDATE
 0 0 /dev/mapper/3600a098038303537762b47594c31586a CANDIDATE
NEWDATA_0003 1 10240 /dev/mapper/3600a098038303537762b47594c315864 MEMBER
NEWDATA_0002 1 10240 /dev/mapper/3600a098038303537762b47594c315863 MEMBER
NEWDATA_0000 1 10240 /dev/mapper/3600a098038303537762b47594c315861 MEMBER
NEWDATA_0001 1 10240 /dev/mapper/3600a098038303537762b47594c315862 MEMBER
```

Adicione novos LUNS

Embora as operações de adição e exclusão possam ser executadas em conjunto, geralmente é mais fácil adicionar novos LUNs em duas etapas. Primeiro, adicione os novos LUNs ao grupo de discos. Essa etapa faz com que metade das extensões sejam migradas dos LUNs ASM atuais para os novos LUNs.

A energia de reequilíbrio indica a taxa à qual os dados estão sendo transferidos. Quanto maior o número, maior o paralelismo da transferência de dados. A migração é realizada com operações de e/S sequenciais eficientes que provavelmente causarão problemas de performance. No entanto, se desejado, o poder de reequilíbrio de uma migração contínua pode ser ajustado com o `alter diskgroup [name] rebalance power [level]` comando. Migrações típicas usam um valor de 5.

```
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586b' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315869' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c315858' rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA add disk
'/dev/mapper/3600a098038303537762b47594c31586a' rebalance power 5;
Diskgroup altered.
```

Monitorização da operação

Uma operação de rebalanceamento pode ser monitorada e gerenciada de várias maneiras. Usamos o seguinte comando para este exemplo.

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

Quando a migração estiver concluída, nenhuma operação de rebalanceamento será relatada.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Soltar LUNs antigos

A migração está agora a meio caminho. Pode ser desejável realizar alguns testes básicos de desempenho para garantir que o ambiente esteja saudável. Após a confirmação, os dados restantes podem ser relocados deixando cair os LUNs antigos. Observe que isso não resulta no lançamento imediato dos LUNs. A operação de queda sinaliza ao Oracle ASM para realocar as extensões primeiro e, em seguida, liberar o LUN.

```
sqlplus / as sysasm
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0000 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup NEWDATA drop disk NEWDATA_0001 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0002 rebalance power 5;
Diskgroup altered.
SQL> alter diskgroup newdata drop disk NEWDATA_0003 rebalance power 5;
Diskgroup altered.
```

Monitorização da operação

A operação de rebalanceamento pode ser monitorada e gerenciada de várias maneiras. Usamos o seguinte comando para este exemplo:

```
SQL> select group_number,operation,state from v$asm_operation;
GROUP_NUMBER OPERA STAT
------------ ----- ----
           1 REBAL RUN
           1 REBAL WAIT
```

Quando a migração estiver concluída, nenhuma operação de rebalanceamento será relatada.

```
SQL> select group_number,operation,state from v$asm_operation;
no rows selected
```

Remova LUNs antigos

Antes de remover os LUNs antigos do grupo de discos, deve efetuar uma verificação final sobre o estado do cabeçalho. Depois que um LUN é liberado do ASM, ele não tem mais um nome listado e o status do cabeçalho é listado como `FORMER`. Isso indica que esses LUNs podem ser removidos com segurança do sistema.

```
SQL> select name||' '||group_number||' '||total_mb||' '||path||'
'||header_status from v$asm_disk;
NAME||''||GROUP_NUMBER||''||TOTAL_MB||''||PATH||''||HEADER_STATUS
--------------------------------------------------------------------------
------
 0 0 /dev/mapper/3600a098038303537762b47594c315863 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315864 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315861 FORMER
 0 0 /dev/mapper/3600a098038303537762b47594c315862 FORMER
NEWDATA_0005 1 10240 /dev/mapper/3600a098038303537762b47594c315869 MEMBER
NEWDATA_0007 1 10240 /dev/mapper/3600a098038303537762b47594c31586a MEMBER
NEWDATA_0004 1 10240 /dev/mapper/3600a098038303537762b47594c31586b MEMBER
NEWDATA_0006 1 10240 /dev/mapper/3600a098038303537762b47594c315858 MEMBER
8 rows selected.
```

Migração para LVM

O procedimento apresentado aqui mostra os princípios de uma migração baseada em LVM de um grupo de volumes `datavg` chamado . Os exemplos são extraídos do LVM Linux, mas os princípios se aplicam igualmente a AIX, HP-UX e VxVM. Os comandos precisos podem variar.

1. Identificar os LUNs atualmente no `datavg` grupo de volumes.

```
[root@host1 ~]# pvdisplay -C | grep datavg
  /dev/mapper/3600a098038303537762b47594c31582f datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c31585a datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c315859 datavg lvm2 a--  10.00g
10.00g
  /dev/mapper/3600a098038303537762b47594c31586c datavg lvm2 a--  10.00g
10.00g
```

2. Crie novos LUNs do mesmo tamanho físico ou ligeiramente maior e defina-os como volumes físicos.

```
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315864
  Physical volume "/dev/mapper/3600a098038303537762b47594c315864"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315863
  Physical volume "/dev/mapper/3600a098038303537762b47594c315863"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315862
  Physical volume "/dev/mapper/3600a098038303537762b47594c315862"
successfully created
[root@host1 ~]# pvcreate /dev/mapper/3600a098038303537762b47594c315861
  Physical volume "/dev/mapper/3600a098038303537762b47594c315861"
successfully created
```

3. Adicione os novos volumes ao grupo de volumes.

```
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315864
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315863
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315862
  Volume group "datavg" successfully extended
[root@host1 tmp]# vgextend datavg
/dev/mapper/3600a098038303537762b47594c315861
  Volume group "datavg" successfully extended
```

4. Emita o `pvmove` comando para realocar as extensões de cada LUN atual para o novo LUN. O `-i [seconds]` argumento monitora o progresso da operação.

```
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31582f
/dev/mapper/3600a098038303537762b47594c315864
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 14.2%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 28.4%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 42.5%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 57.1%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 72.3%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 87.3%
  /dev/mapper/3600a098038303537762b47594c31582f: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31585a
/dev/mapper/3600a098038303537762b47594c315863
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 14.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 29.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 44.8%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 60.1%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 75.8%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 90.9%
  /dev/mapper/3600a098038303537762b47594c31585a: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c315859
/dev/mapper/3600a098038303537762b47594c315862
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 14.8%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 29.8%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 45.5%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 61.1%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 76.6%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 91.7%
  /dev/mapper/3600a098038303537762b47594c315859: Moved: 100.0%
[root@host1 tmp]# pvmove -i 10
/dev/mapper/3600a098038303537762b47594c31586c
/dev/mapper/3600a098038303537762b47594c315861
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 0.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 15.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 30.4%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 46.0%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 61.4%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 77.2%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 92.3%
  /dev/mapper/3600a098038303537762b47594c31586c: Moved: 100.0%
```

5. Quando esse processo estiver concluído, solte os LUNs antigos do grupo de volumes usando o `vgreduce` comando. Se for bem-sucedido, o LUN pode agora ser removido de forma segura do sistema.

```
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31582f
Removed "/dev/mapper/3600a098038303537762b47594c31582f" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31585a
  Removed "/dev/mapper/3600a098038303537762b47594c31585a" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c315859
  Removed "/dev/mapper/3600a098038303537762b47594c315859" from volume
group "datavg"
[root@host1 tmp]# vgreduce datavg
/dev/mapper/3600a098038303537762b47594c31586c
  Removed "/dev/mapper/3600a098038303537762b47594c31586c" from volume
group "datavg"
```

Importação LUN estrangeiro

Planejamento

Os procedimentos para migrar recursos SAN usando FLI estão documentados no NetApp "Documentação de importação de LUN estrangeiro da ONTAP" .

Do ponto de vista de um banco de dados e host, não são necessárias etapas especiais. Depois que as zonas FC forem atualizadas e os LUNs ficarem disponíveis no ONTAP, o LVM poderá ler os metadados do LVM dos LUNs. Além disso, os grupos de volume estão prontos para uso sem etapas de configuração adicionais. Em casos raros, os ambientes podem incluir arquivos de configuração que foram codificados com referências à matriz de armazenamento anterior. Por exemplo, um sistema Linux que incluía `/etc/multipath.conf` regras que referiam uma WWN de um determinado dispositivo deve ser atualizado para refletir as alterações introduzidas pela FLI.

> ⓘ Consulte a Matriz de compatibilidade do NetApp para obter informações sobre as configurações suportadas. Se o seu ambiente não estiver incluído, contacte o seu representante da NetApp para obter assistência.

Este exemplo mostra a migração de LUNs ASM e LVM hospedados em um servidor Linux. O FLI é suportado em outros sistemas operacionais e, embora os comandos do lado do host possam diferir, os princípios são os mesmos e os procedimentos do ONTAP são idênticos.

Identificar LUNs LVM

O primeiro passo em preparação é identificar os LUNs a serem migrados. No exemplo mostrado aqui, dois sistemas de arquivos baseados em SAN são montados `/orabin` em e `/backups`.

```
[root@host1 ~]# df -k
Filesystem                          1K-blocks      Used Available Use%
Mounted on
/dev/mapper/rhel-root                52403200   8811464  43591736  17% /
devtmpfs                             65882776         0  65882776   0% /dev
...
fas8060-nfs-public:/install         199229440 119368128  79861312  60%
/install
/dev/mapper/sanvg-lvorabin           20961280  12348476   8612804  59%
/orabin
/dev/mapper/sanvg-lvbackups          73364480  62947536  10416944  86%
/backups
```

O nome do grupo de volume pode ser extraído do nome do dispositivo, que usa o formato (nome do grupo de volume)-(nome do volume lógico). Neste caso, o grupo de volume é `sanvg` chamado .

O `pvdisplay` comando pode ser usado da seguinte forma para identificar os LUNs que suportam este grupo de volumes. Nesse caso, existem 10 LUNs que compõem o `sanvg` grupo de volumes.

```
[root@host1 ~]# pvdisplay -C -o pv_name,pv_size,pv_fmt,vg_name
  PV                                             PSize   VG
  /dev/mapper/3600a0980383030445424487556574266  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574267  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574268  10.00g sanvg
  /dev/mapper/3600a0980383030445424487556574269  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426a  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426b  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426c  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426d  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426e  10.00g sanvg
  /dev/mapper/3600a098038303044542448755657426f  10.00g sanvg
  /dev/sda2                                     278.38g rhel
```

Identificar LUNs ASM

LUNs ASM também devem ser migrados. Para obter o número de LUNs e caminhos LUN do sqlplus como usuário do sysasm, execute o seguinte comando:

```
SQL> select path||' '||os_mb from v$asm_disk;
PATH||''||OS_MB
--------------------------------------------------------------------------
------
/dev/oracleasm/disks/ASM0 10240
/dev/oracleasm/disks/ASM9 10240
/dev/oracleasm/disks/ASM8 10240
/dev/oracleasm/disks/ASM7 10240
/dev/oracleasm/disks/ASM6 10240
/dev/oracleasm/disks/ASM5 10240
/dev/oracleasm/disks/ASM4 10240
/dev/oracleasm/disks/ASM1 10240
/dev/oracleasm/disks/ASM3 10240
/dev/oracleasm/disks/ASM2 10240
10 rows selected.
SQL>
```

Alterações na rede FC

O ambiente atual contém 20 LUNs a serem migrados. Atualize a SAN atual para que o ONTAP possa acessar os LUNs atuais. Os dados ainda não foram migrados, mas o ONTAP deve ler informações de configuração dos LUNs atuais para criar a nova casa para esses dados.

No mínimo, pelo menos uma porta HBA no sistema AFF/FAS deve ser configurada como uma porta do iniciador. Além disso, as zonas FC precisam ser atualizadas para que o ONTAP possa acessar os LUNs no storage array estrangeiro. Alguns storages de armazenamento têm o mascaramento LUN configurado, o que limita quais WWNs podem acessar um determinado LUN. Nesses casos, o mascaramento de LUN também deve ser atualizado para conceder acesso às WWNs do ONTAP.

Depois que esta etapa for concluída, o ONTAP deve ser capaz de visualizar a matriz de armazenamento estrangeira com o `storage array show` comando. O campo chave que retorna é o prefixo usado para identificar o LUN estranho no sistema. No exemplo abaixo, os LUNs na matriz estrangeira `FOREIGN_1` aparecem no ONTAP usando o prefixo do `FOR-1`.

Identificar matriz estrangeira

```
Cluster01::> storage array show -fields name,prefix
name          prefix
------------- ------
FOREIGN_1     FOR-1
Cluster01::>
```

Identificar LUNs estranhos

Os LUNs podem ser listados passando o `array-name` para o `storage disk show` comando. Os dados retornados são referenciados várias vezes durante o procedimento de migração.

```
Cluster01::> storage disk show -array-name FOREIGN_1 -fields disk,serial
disk     serial-number
-------- -------------
FOR-1.1  800DT$HuVWBX
FOR-1.2  800DT$HuVWBZ
FOR-1.3  800DT$HuVWBW
FOR-1.4  800DT$HuVWBY
FOR-1.5  800DT$HuVWB/
FOR-1.6  800DT$HuVWBa
FOR-1.7  800DT$HuVWBd
FOR-1.8  800DT$HuVWBb
FOR-1.9  800DT$HuVWBc
FOR-1.10 800DT$HuVWBe
FOR-1.11 800DT$HuVWBf
FOR-1.12 800DT$HuVWBg
FOR-1.13 800DT$HuVWBi
FOR-1.14 800DT$HuVWBh
FOR-1.15 800DT$HuVWBj
FOR-1.16 800DT$HuVWBk
FOR-1.17 800DT$HuVWBm
FOR-1.18 800DT$HuVWBl
FOR-1.19 800DT$HuVWBo
FOR-1.20 800DT$HuVWBn
20 entries were displayed.
Cluster01::>
```

Registre LUNs de matriz estrangeira como candidatos à importação

Os LUNs estrangeiros são inicialmente classificados como qualquer tipo de LUN específico. Antes que os dados possam ser importados, os LUNs devem ser marcados como estrangeiros e, portanto, um candidato para o processo de importação. Esta etapa é concluída passando o número de série para `storage disk modify` o comando, como mostrado no exemplo a seguir. Observe que esse processo marca somente o LUN como estranho dentro do ONTAP. Nenhum dado é gravado no próprio LUN estrangeiro.

```
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBW} -is
-foreign true
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBX} -is
-foreign true
...
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBn} -is
-foreign true
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBo} -is
-foreign true
Cluster01::*>
```

Criar volumes para hospedar LUNs migrados

É necessário um volume para hospedar os LUNs migrados. A configuração exata do volume depende do plano geral para aproveitar os recursos do ONTAP. Neste exemplo, os LUNs ASM são colocados em um volume e os LUNs LVM são colocados em um segundo volume. Com isso, você pode gerenciar os LUNs como grupos independentes para fins como disposição em camadas, criação de snapshots ou configuração de controles de QoS.

Defina o `snapshot-policy `to `none`. O processo de migração pode incluir uma grande quantidade de rotatividade de dados. Portanto, pode haver um grande aumento no consumo de espaço se os snapshots forem criados acidentalmente porque os dados indesejados são capturados nos snapshots.

```
Cluster01::> volume create -volume new_asm -aggregate data_02 -size 120G
-snapshot-policy none
[Job 1152] Job succeeded: Successful
Cluster01::> volume create -volume new_lvm -aggregate data_02 -size 120G
-snapshot-policy none
[Job 1153] Job succeeded: Successful
Cluster01::>
```

Criar LUNs ONTAP

Após a criação dos volumes, é necessário criar os novos LUNs. Normalmente, a criação de um LUN requer que o usuário especifique tais informações como o tamanho do LUN, mas neste caso o argumento de disco externo é passado para o comando. Como resultado, o ONTAP replica os dados de configuração de LUN atuais a partir do número de série especificado. Ele também usa a geometria LUN e os dados da tabela de partição para ajustar o alinhamento LUN e estabelecer o desempenho ideal.

Nesta etapa, os números de série devem ser cruzados em relação à matriz estrangeira para garantir que o LUN estranho correto seja correspondido ao novo LUN correto.

```
Cluster01::*> lun create -vserver vserver1 -path /vol/new_asm/LUN0 -ostype
linux -foreign-disk 800DT$HuVWBW
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_asm/LUN1 -ostype
linux -foreign-disk 800DT$HuVWBX
Created a LUN of size 10g (10737418240)
...
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_lvm/LUN8 -ostype
linux -foreign-disk 800DT$HuVWBn
Created a LUN of size 10g (10737418240)
Cluster01::*> lun create -vserver vserver1 -path /vol/new_lvm/LUN9 -ostype
linux -foreign-disk 800DT$HuVWBo
Created a LUN of size 10g (10737418240)
```

Crie relações de importação

Os LUNs agora foram criados, mas não estão configurados como um destino de replicação. Antes que essa etapa possa ser executada, os LUNs devem primeiro ser colocados off-line. Esta etapa extra foi projetada para proteger dados contra erros do usuário. Se o ONTAP permitisse que uma migração fosse executada em um LUN on-line, isso criaria o risco de que um erro tipográfico pudesse resultar na substituição de dados ativos. A etapa adicional de forçar o usuário a primeiro colocar um LUN off-line ajuda a verificar se o LUN de destino correto é usado como um destino de migração.

```
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_asm/LUN0
Warning: This command will take LUN "/vol/new_asm/LUN0" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_asm/LUN1
Warning: This command will take LUN "/vol/new_asm/LUN1" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
...
Warning: This command will take LUN "/vol/new_lvm/LUN8" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
Cluster01::*> lun offline -vserver vserver1 -path /vol/new_lvm/LUN9
Warning: This command will take LUN "/vol/new_lvm/LUN9" in Vserver
         "vserver1" offline.
Do you want to continue? {y|n}: y
```

Depois que os LUNs estiverem offline, você pode estabelecer a relação de importação passando o número de série LUN estrangeiro para `lun import create` o comando.

```
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_asm/LUN0
-foreign-disk 800DT$HuVWBW
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_asm/LUN1
-foreign-disk 800DT$HuVWBX
...
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_lvm/LUN8
-foreign-disk 800DT$HuVWBn
Cluster01::*> lun import create -vserver vserver1 -path /vol/new_lvm/LUN9
-foreign-disk 800DT$HuVWBo
Cluster01::*>
```

Depois que todas as relações de importação forem estabelecidas, os LUNs podem ser colocados online novamente.

```
Cluster01::*> lun online -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun online -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun online -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun online -vserver vserver1 -path /vol/new_lvm/LUN9
Cluster01::*>
```

Criar grupo de iniciadores

Um grupo de iniciadores (igroup) faz parte da arquitetura de mascaramento de LUN do ONTAP. Um LUN recém-criado não é acessível a menos que um host tenha acesso concedido pela primeira vez. Isso é feito criando um grupo que lista os nomes dos iniciadores FC WWNs ou iSCSI que devem ser concedidos acesso. Na época em que esse relatório foi escrito, a FLI era compatível apenas com LUNs FC. No entanto, a conversão para iSCSI pós-migração é uma tarefa simples, como mostrado na "Conversão de protocolo".

Neste exemplo, um grupo é criado que contém duas WWNs que correspondem às duas portas disponíveis no HBA do host.

```
Cluster01::*> igroup create linuxhost -protocol fcp -ostype linux
-initiator 21:00:00:0e:1e:16:63:50 21:00:00:0e:1e:16:63:51
```

Mapear novos LUNs para o host

Após a criação do grupo, os LUNs são então mapeados para o grupo definido. Esses LUNs estão disponíveis apenas para as WWNs incluídas neste grupo. O NetApp assume nesta fase do processo de migração que o host não foi zoneado para o ONTAP. Isso é importante porque, se o host for simultaneamente zoneado para o array estrangeiro e o novo sistema ONTAP, existe o risco de que LUNs com o mesmo número de série possam ser descobertos em cada array. Essa situação pode levar a falhas de multipath ou danos aos dados.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxhost
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxhost
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxhost
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxhost
Cluster01::*>
```

Redução

Alguma interrupção durante uma importação LUN estrangeira é inevitável devido à necessidade de alterar a configuração da rede FC. No entanto, a interrupção não precisa durar muito mais do que o tempo necessário para reiniciar o ambiente de banco de

dados e atualizar o zoneamento FC para alternar a conetividade FC do host do LUN externo para o ONTAP.

Este processo pode ser resumido da seguinte forma:

1. Quiesce toda a atividade LUN nos LUNs externos.
2. Redirecione as conexões FC do host para o novo sistema ONTAP.
3. Acione o processo de importação.
4. Redescubra os LUNs.
5. Reinicie o banco de dados.

Não é necessário esperar que o processo de migração seja concluído. Assim que a migração para um determinado LUN começar, ele estará disponível no ONTAP e poderá servir dados enquanto o processo de cópia de dados continuar. Todas as leituras são passadas para o LUN estrangeiro, e todas as gravações são escritas de forma síncrona em ambos os arrays. A operação de cópia é muito rápida e a sobrecarga de redirecionar o tráfego FC é mínima, portanto, qualquer impactos no desempenho deve ser transitório e mínimo. Se houver problema, você pode atrasar a reinicialização do ambiente até que o processo de migração seja concluído e as relações de importação tenham sido excluídas.

Encerre a base de dados

O primeiro passo para silenciar o ambiente neste exemplo é desligar o banco de dados.

```
[oracle@host1 bin]$ . oraenv
ORACLE_SID = [oracle] ? FLIDB
The Oracle base remains unchanged with value /orabin
[oracle@host1 bin]$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to:
Oracle Database 12c Enterprise Edition Release 12.1.0.2.0 - 64bit
Production
With the Partitioning, Automatic Storage Management, OLAP, Advanced
Analytics
and Real Application Testing options
SQL> shutdown immediate;
Database closed.
Database dismounted.
ORACLE instance shut down.
SQL>
```

Encerre os serviços da grade

Um dos sistemas de arquivos baseados em SAN que está sendo migrado também inclui os serviços Oracle ASM. A supressão dos LUNs subjacentes requer a desmontagem dos sistemas de arquivos, o que, por sua vez, significa parar todos os processos com arquivos abertos neste sistema de arquivos.

```
[oracle@host1 bin]$ ./crsctl stop has -f
CRS-2791: Starting shutdown of Oracle High Availability Services-managed
resources on 'host1'
CRS-2673: Attempting to stop 'ora.evmd' on 'host1'
CRS-2673: Attempting to stop 'ora.DATA.dg' on 'host1'
CRS-2673: Attempting to stop 'ora.LISTENER.lsnr' on 'host1'
CRS-2677: Stop of 'ora.DATA.dg' on 'host1' succeeded
CRS-2673: Attempting to stop 'ora.asm' on 'host1'
CRS-2677: Stop of 'ora.LISTENER.lsnr' on 'host1' succeeded
CRS-2677: Stop of 'ora.evmd' on 'host1' succeeded
CRS-2677: Stop of 'ora.asm' on 'host1' succeeded
CRS-2673: Attempting to stop 'ora.cssd' on 'host1'
CRS-2677: Stop of 'ora.cssd' on 'host1' succeeded
CRS-2793: Shutdown of Oracle High Availability Services-managed resources
on 'host1' has completed
CRS-4133: Oracle High Availability Services has been stopped.
[oracle@host1 bin]$
```

Desmontar sistemas de arquivos

Se todos os processos forem desligados, a operação umount será bem-sucedida. Se a permissão for negada, deve haver um processo com um bloqueio no sistema de arquivos. O `fuser` comando pode ajudar a identificar esses processos.

```
[root@host1 ~]# umount /orabin
[root@host1 ~]# umount /backups
```

Desativar grupos de volume

Depois de todos os sistemas de arquivos em um determinado grupo de volume serem desmontados, o grupo de volume pode ser desativado.

```
[root@host1 ~]# vgchange --activate n sanvg
  0 logical volume(s) in volume group "sanvg" now active
[root@host1 ~]#
```

Alterações na rede FC

As zonas FC agora podem ser atualizadas para remover todo o acesso do host ao array externo e estabelecer acesso ao ONTAP.

Inicie o processo de importação

Para iniciar os processos de importação de LUN, execute o `lun import start` comando.

```
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_asm/LUN0
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_asm/LUN1
...
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_lvm/LUN8
Cluster01::lun import*> lun import start -vserver vserver1 -path
/vol/new_lvm/LUN9
Cluster01::lun import*>
```

Monitorar o progresso da importação

A operação de importação pode ser monitorada com o `lun import show` comando. Como mostrado abaixo, a importação de todos os LUNs 20 está em andamento, o que significa que os dados agora estão acessíveis por meio do ONTAP, mesmo que a operação de cópia de dados ainda progrida.

```
Cluster01::lun import*> lun import show -fields path,percent-complete
vserver   foreign-disk path              percent-complete
--------- ------------ ----------------- ----------------
vserver1  800DT$HuVWB/ /vol/new_asm/LUN4 5
vserver1  800DT$HuVWBW /vol/new_asm/LUN0 5
vserver1  800DT$HuVWBX /vol/new_asm/LUN1 6
vserver1  800DT$HuVWBY /vol/new_asm/LUN2 6
vserver1  800DT$HuVWBZ /vol/new_asm/LUN3 5
vserver1  800DT$HuVWBa /vol/new_asm/LUN5 4
vserver1  800DT$HuVWBb /vol/new_asm/LUN6 4
vserver1  800DT$HuVWBc /vol/new_asm/LUN7 4
vserver1  800DT$HuVWBd /vol/new_asm/LUN8 4
vserver1  800DT$HuVWBe /vol/new_asm/LUN9 4
vserver1  800DT$HuVWBf /vol/new_lvm/LUN0 5
vserver1  800DT$HuVWBg /vol/new_lvm/LUN1 4
vserver1  800DT$HuVWBh /vol/new_lvm/LUN2 4
vserver1  800DT$HuVWBi /vol/new_lvm/LUN3 3
vserver1  800DT$HuVWBj /vol/new_lvm/LUN4 3
vserver1  800DT$HuVWBk /vol/new_lvm/LUN5 3
vserver1  800DT$HuVWBl /vol/new_lvm/LUN6 4
vserver1  800DT$HuVWBm /vol/new_lvm/LUN7 3
vserver1  800DT$HuVWBn /vol/new_lvm/LUN8 2
vserver1  800DT$HuVWBo /vol/new_lvm/LUN9 2
20 entries were displayed.
```

Se você precisar de um processo off-line, atrasar a redescoberta ou a reinicialização dos serviços até que o `lun import show` comando indique que toda a migração foi bem-sucedida e concluída. Em seguida, você pode concluir o processo de migração conforme descrito em "Importação LUN estrangeiro - conclusão".

Se precisar de uma migração online, continue a redescobrir os LUNs na sua nova casa e a abrir os serviços.

Verifique se há alterações no dispositivo SCSI

Na maioria dos casos, a opção mais simples para redescobrir novos LUNs é reiniciar o host. Isso remove automaticamente dispositivos obsoletos antigos, descobre corretamente todos os novos LUNs e cria dispositivos associados, como dispositivos multipathing. O exemplo aqui mostra um processo totalmente online para fins de demonstração.

Cuidado: Antes de reiniciar um host, certifique-se de que todas as entradas `/etc/fstab` nessa referência migradas dos recursos SAN sejam comentadas. Se isso não for feito e houver problemas com o acesso LUN, o sistema operacional pode não inicializar. Esta situação não danifica os dados. No entanto, pode ser muito inconveniente inicializar no modo de recuperação ou em um modo semelhante e corrigir o `/etc/fstab` para que o sistema operacional possa ser inicializado para habilitar a solução de problemas.

Os LUNs na versão do Linux usada neste exemplo podem ser reconfigurados com o `rescan-scsi-bus.sh` comando. Se o comando for bem-sucedido, cada caminho LUN deverá aparecer na saída. A saída pode ser difícil de interpretar, mas, se o zoneamento e a configuração do igrop estavam corretos, muitos LUNs devem aparecer que incluem uma `NETAPP` string de fornecedor.

```
[root@host1 /]# rescan-scsi-bus.sh
Scanning SCSI subsystem for new devices
Scanning host 0 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
 Scanning for device 0 2 0 0 ...
OLD: Host: scsi0 Channel: 02 Id: 00 Lun: 00
      Vendor: LSI      Model: RAID SAS 6G 0/1  Rev: 2.13
      Type:   Direct-Access                    ANSI SCSI revision: 05
Scanning host 1 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
 Scanning for device 1 0 0 0 ...
OLD: Host: scsi1 Channel: 00 Id: 00 Lun: 00
      Vendor: Optiarc  Model: DVD RW AD-7760H  Rev: 1.41
      Type:   CD-ROM                           ANSI SCSI revision: 05
Scanning host 2 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 3 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 4 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 5 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 6 for  SCSI target IDs  0 1 2 3 4 5 6 7, all LUNs
Scanning host 7 for  all SCSI target IDs, all LUNs
 Scanning for device 7 0 0 10 ...
OLD: Host: scsi7 Channel: 00 Id: 00 Lun: 10
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 7 0 0 11 ...
OLD: Host: scsi7 Channel: 00 Id: 00 Lun: 11
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 7 0 0 12 ...
...
OLD: Host: scsi9 Channel: 00 Id: 01 Lun: 18
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
 Scanning for device 9 0 1 19 ...
OLD: Host: scsi9 Channel: 00 Id: 01 Lun: 19
      Vendor: NETAPP   Model: LUN C-Mode       Rev: 8300
      Type:   Direct-Access                    ANSI SCSI revision: 05
0 new or changed device(s) found.
0 remapped or resized device(s) found.
0 device(s) removed.
```

Verifique se existem dispositivos multipath

O processo de descoberta LUN também aciona a recriação de dispositivos multipath, mas o driver de multipathing Linux é conhecido por ter problemas ocasionais. A saída de `multipath -ll` deve ser verificada para verificar se a saída parece como esperado. Por exemplo, a saída abaixo mostra os dispositivos multipath associados a uma `NETAPP` cadeia de carateres de fornecedor. Cada dispositivo tem quatro caminhos, com dois em uma prioridade de 50 e dois em uma prioridade de 10. Embora a saída exata possa

variar com diferentes versões do Linux, essa saída parece como esperado.

Consulte a documentação dos utilitários do host para a versão do Linux que você usa para verificar se as `/etc/multipath.conf` configurações estão corretas.

```
[root@host1 /]# multipath -ll
3600a098038303558735d493762504b36 dm-5 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:4  sdat 66:208 active ready running
| `- 9:0:1:4  sdbn 68:16  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:4  sdf  8:80   active ready running
  `- 9:0:0:4  sdz  65:144 active ready running
3600a098038303558735d493762504b2d dm-10 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:8  sdax 67:16  active ready running
| `- 9:0:1:8  sdbr 68:80  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:8  sdj  8:144  active ready running
  `- 9:0:0:8  sdad 65:208 active ready running
...
3600a098038303558735d493762504b37 dm-8 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:5  sdau 66:224 active ready running
| `- 9:0:1:5  sdbo 68:32  active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:5  sdg  8:96   active ready running
  `- 9:0:0:5  sdaa 65:160 active ready running
3600a098038303558735d493762504b4b dm-22 NETAPP  ,LUN C-Mode
size=10G features='4 queue_if_no_path pg_init_retries 50
retain_attached_hw_handle' hwhandler='1 alua' wp=rw
|-+- policy='service-time 0' prio=50 status=active
| |- 7:0:1:19 sdbi 67:192 active ready running
| `- 9:0:1:19 sdcc 69:0   active ready running
`-+- policy='service-time 0' prio=10 status=enabled
  |- 7:0:0:19 sdu  65:64  active ready running
  `- 9:0:0:19 sdao 66:128 active ready running
```

Reative o grupo de volumes LVM

Se os LUNs LVM tiverem sido detetados corretamente, o `vgchange --activate y` comando deverá ser bem-sucedido. Este é um bom exemplo do valor de um gerenciador de volume lógico. Uma alteração na WWN de um LUN ou mesmo de um número de série não é importante porque os metadados do grupo de volume são gravados no próprio LUN.

O sistema operacional digitalizou os LUNs e descobriu uma pequena quantidade de dados gravados no LUN que os identifica como um volume físico pertencente ao `sanvg volumegroup`. Em seguida, ele construiu todos os dispositivos necessários. Tudo o que é necessário é reativar o grupo de volume.

```
[root@host1 /]# vgchange --activate y sanvg
  Found duplicate PV fpCzdLTuKfy2xDZjai1NliJh3TjLUBiT: using
/dev/mapper/3600a098038303558735d493762504b46 not /dev/sdp
  Using duplicate PV /dev/mapper/3600a098038303558735d493762504b46 from
subsystem DM, ignoring /dev/sdp
  2 logical volume(s) in volume group "sanvg" now active
```

Remontagem dos sistemas de arquivos

Depois que o grupo de volume é reativado, os sistemas de arquivos podem ser montados com todos os dados originais intactos. Como discutido anteriormente, os sistemas de arquivos estão totalmente operacionais, mesmo que a replicação de dados ainda esteja ativa no grupo de volta.

```
[root@host1 /]# mount /orabin
[root@host1 /]# mount /backups
[root@host1 /]# df -k
Filesystem                          1K-blocks      Used Available Use%
Mounted on
/dev/mapper/rhel-root                52403200   8837100  43566100  17% /
devtmpfs                             65882776         0  65882776   0% /dev
tmpfs                                 6291456        84   6291372   1%
/dev/shm
tmpfs                                65898668      9884  65888784   1% /run
tmpfs                                65898668         0  65898668   0%
/sys/fs/cgroup
/dev/sda1                              505580    224828    280752  45% /boot
fas8060-nfs-public:/install          199229440 119368256  79861184  60%
/install
fas8040-nfs-routable:/snapomatic      9961472     30528   9930944   1%
/snapomatic
tmpfs                                13179736        16  13179720   1%
/run/user/42
tmpfs                                13179736         0  13179736   0%
/run/user/0
/dev/mapper/sanvg-lvorabin           20961280  12357456   8603824  59%
/orabin
/dev/mapper/sanvg-lvbackups          73364480  62947536  10416944  86%
/backups
```

Redigitalização para dispositivos ASM

Os dispositivos ASMlib devem ter sido redescobertos quando os dispositivos SCSI foram reconfigurados. A redescoberta pode ser verificada on-line reiniciando o ASMlib e, em seguida, digitalizando os discos.

Esta etapa só é relevante para configurações ASM onde ASMlib é usado.

Atenção: Onde o ASMlib não é usado, os `/dev/mapper` dispositivos devem ter sido recriados automaticamente. No entanto, as permissões podem não estar corretas. Você deve definir permissões especiais nos dispositivos subjacentes para ASM na ausência de ASMlib. Isso geralmente é feito através de entradas especiais nas `/etc/multipath.conf` regras ou `udev`, ou possivelmente em ambos os conjuntos de regras. Esses arquivos podem precisar ser atualizados para refletir alterações no ambiente em termos de WWNs ou números de série para garantir que os dispositivos ASM ainda tenham as permissões corretas.

Neste exemplo, reiniciar o ASMlib e procurar discos mostra os mesmos LUNs ASM 10 do ambiente original.

```
[root@host1 /]# oracleasm exit
Unmounting ASMlib driver filesystem: /dev/oracleasm
Unloading module "oracleasm": oracleasm
[root@host1 /]# oracleasm init
Loading module "oracleasm": oracleasm
Configuring "oracleasm" to use device physical block size
Mounting ASMlib driver filesystem: /dev/oracleasm
[root@host1 /]# oracleasm scandisks
Reloading disk partitions: done
Cleaning any stale ASM disks...
Scanning system for ASM disks...
Instantiating disk "ASM0"
Instantiating disk "ASM1"
Instantiating disk "ASM2"
Instantiating disk "ASM3"
Instantiating disk "ASM4"
Instantiating disk "ASM5"
Instantiating disk "ASM6"
Instantiating disk "ASM7"
Instantiating disk "ASM8"
Instantiating disk "ASM9"
```

Reinicie os serviços de grade

Agora que os dispositivos LVM e ASM estão online e disponíveis, os serviços de grade podem ser reiniciados.

```
[root@host1 /]# cd /orabin/product/12.1.0/grid/bin
[root@host1 bin]# ./crsctl start has
```

Reinicie a base de dados

Depois que os serviços de grade tiverem sido reiniciados, o banco de dados pode ser criado. Pode ser necessário esperar alguns minutos para que os serviços ASM fiquem totalmente disponíveis antes de tentar iniciar o banco de dados.

```
[root@host1 bin]# su - oracle
[oracle@host1 ~]$ . oraenv
ORACLE_SID = [oracle] ? FLIDB
The Oracle base has been set to /orabin
[oracle@host1 ~]$ sqlplus / as sysdba
SQL*Plus: Release 12.1.0.2.0
Copyright (c) 1982, 2014, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> startup
ORACLE instance started.
Total System Global Area 3221225472 bytes
Fixed Size                  4502416 bytes
Variable Size            1207962736 bytes
Database Buffers         1996488704 bytes
Redo Buffers               12271616 bytes
Database mounted.
Database opened.
SQL>
```

Conclusão

Do ponto de vista do host, a migração está concluída, mas a e/S ainda é atendida do array estrangeiro até que as relações de importação sejam excluídas.

Antes de excluir os relacionamentos, confirme se o processo de migração está concluído para todos os LUNs.

```
Cluster01::*> lun import show -vserver vserver1 -fields foreign-
disk,path,operational-state
vserver   foreign-disk path              operational-state
--------- ------------ ----------------- -----------------
vserver1 800DT$HuVWB/ /vol/new_asm/LUN4 completed
vserver1 800DT$HuVWBW /vol/new_asm/LUN0 completed
vserver1 800DT$HuVWBX /vol/new_asm/LUN1 completed
vserver1 800DT$HuVWBY /vol/new_asm/LUN2 completed
vserver1 800DT$HuVWBZ /vol/new_asm/LUN3 completed
vserver1 800DT$HuVWBa /vol/new_asm/LUN5 completed
vserver1 800DT$HuVWBb /vol/new_asm/LUN6 completed
vserver1 800DT$HuVWBc /vol/new_asm/LUN7 completed
vserver1 800DT$HuVWBd /vol/new_asm/LUN8 completed
vserver1 800DT$HuVWBe /vol/new_asm/LUN9 completed
vserver1 800DT$HuVWBf /vol/new_lvm/LUN0 completed
vserver1 800DT$HuVWBg /vol/new_lvm/LUN1 completed
vserver1 800DT$HuVWBh /vol/new_lvm/LUN2 completed
vserver1 800DT$HuVWBi /vol/new_lvm/LUN3 completed
vserver1 800DT$HuVWBj /vol/new_lvm/LUN4 completed
vserver1 800DT$HuVWBk /vol/new_lvm/LUN5 completed
vserver1 800DT$HuVWBl /vol/new_lvm/LUN6 completed
vserver1 800DT$HuVWBm /vol/new_lvm/LUN7 completed
vserver1 800DT$HuVWBn /vol/new_lvm/LUN8 completed
vserver1 800DT$HuVWBo /vol/new_lvm/LUN9 completed
20 entries were displayed.
```

Eliminar relações de importação

Quando o processo de migração estiver concluído, exclua a relação de migração. Depois de fazer isso, e/S é servido exclusivamente a partir das unidades no ONTAP.

```
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN0
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_asm/LUN1
...
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN8
Cluster01::*> lun import delete -vserver vserver1 -path /vol/new_lvm/LUN9
```

Anular o registo de LUNs estranhos

Finalmente, modifique o disco para remover a `is-foreign` designação.

```
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBW} -is
-foreign false
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBX} -is
-foreign false
...
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBn} -is
-foreign false
Cluster01::*> storage disk modify {-serial-number 800DT$HuVWBo} -is
-foreign false
Cluster01::*>
```

Conversão de protocolo

Alterar o protocolo usado para acessar um LUN é um requisito comum.

Em alguns casos, faz parte de uma estratégia geral para migrar dados para a nuvem. O TCP/IP é o protocolo da nuvem e a mudança de FC para iSCSI permite uma migração mais fácil para vários ambientes de nuvem. Em outros casos, o iSCSI pode ser desejável para aproveitar os custos reduzidos de uma SAN IP. Ocasionalmente, uma migração pode usar um protocolo diferente como medida temporária. Por exemplo, se um array estrangeiro e LUNs baseados em ONTAP não contiverem nos mesmos HBAs, você poderá usar LUNs iSCSI o suficiente para copiar dados do array antigo. Em seguida, você pode converter de volta para FC depois que os LUNs antigos forem removidos do sistema.

O procedimento a seguir demonstra a conversão de FC para iSCSI, mas os princípios gerais se aplicam a uma conversão de iSCSI para FC reversa.

Instale o iniciador iSCSI

A maioria dos sistemas operacionais inclui um iniciador iSCSI de software por padrão, mas se um não estiver incluído, ele pode ser facilmente instalado.

```
[root@host1 /]# yum install -y iscsi-initiator-utils
Loaded plugins: langpacks, product-id, search-disabled-repos,
subscription-
              : manager
Resolving Dependencies
--> Running transaction check
---> Package iscsi-initiator-utils.x86_64 0:6.2.0.873-32.el7 will be
updated
--> Processing Dependency: iscsi-initiator-utils = 6.2.0.873-32.el7 for
package: iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
---> Package iscsi-initiator-utils.x86_64 0:6.2.0.873-32.0.2.el7 will be
an update
--> Running transaction check
---> Package iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.el7 will
be updated
---> Package iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.0.2.el7
```

```
will be an update
--> Finished Dependency Resolution
Dependencies Resolved
================================================================================
===
 Package                         Arch   Version              Repository
Size
================================================================================
===
Updating:
 iscsi-initiator-utils           x86_64 6.2.0.873-32.0.2.el7 ol7_latest 416
k
Updating for dependencies:
 iscsi-initiator-utils-iscsiuio x86_64 6.2.0.873-32.0.2.el7 ol7_latest  84
k
Transaction Summary
================================================================================
===
Upgrade  1 Package (+1 Dependent package)
Total download size: 501 k
Downloading packages:
No Presto metadata available for ol7_latest
(1/2): iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_6 | 416 kB   00:00
(2/2): iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2. |  84 kB   00:00
--------------------------------------------------------------------------------
---
Total                                           2.8 MB/s | 501 kB
00:00Cluster01
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
  Updating   : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2.el7.x86
1/4
  Updating   : iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_64
2/4
  Cleanup    : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
3/4
  Cleanup    : iscsi-initiator-utils-6.2.0.873-32.el7.x86_64
4/4
rhel-7-server-eus-rpms/7Server/x86_64/productid       | 1.7 kB     00:00
rhel-7-server-rpms/7Server/x86_64/productid           | 1.7 kB     00:00
  Verifying  : iscsi-initiator-utils-6.2.0.873-32.0.2.el7.x86_64
1/4
  Verifying  : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.0.2.el7.x86
2/4
```

```
  Verifying  : iscsi-initiator-utils-iscsiuio-6.2.0.873-32.el7.x86_64
3/4
  Verifying  : iscsi-initiator-utils-6.2.0.873-32.el7.x86_64
4/4
Updated:
  iscsi-initiator-utils.x86_64 0:6.2.0.873-32.0.2.el7
Dependency Updated:
  iscsi-initiator-utils-iscsiuio.x86_64 0:6.2.0.873-32.0.2.el7
Complete!
[root@host1 /]#
```

Identificar o nome do iniciador iSCSI

Um nome exclusivo do iniciador iSCSI é gerado durante o processo de instalação. No Linux, ele está localizado no `/etc/iscsi/initiatorname.iscsi` arquivo. Esse nome é usado para identificar o host na SAN IP.

```
[root@host1 /]# cat /etc/iscsi/initiatorname.iscsi
InitiatorName=iqn.1992-05.com.redhat:497bd66ca0
```

Crie um novo grupo de iniciadores

Um grupo de iniciadores (igroup) faz parte da arquitetura de mascaramento de LUN do ONTAP. Um LUN recém-criado não é acessível a menos que um host tenha acesso concedido pela primeira vez. Esta etapa é realizada criando um grupo que lista os nomes de iniciador iSCSI ou WWNs FC que exigem acesso.

Neste exemplo, um grupo é criado que contém o iniciador iSCSI do host Linux.

```
Cluster01::*> igroup create -igroup linuxiscsi -protocol iscsi -ostype
linux -initiator iqn.1994-05.com.redhat:497bd66ca0
```

Encerrar o ambiente

Antes de alterar o protocolo LUN, os LUNs devem estar totalmente quietos. Qualquer banco de dados em um dos LUNs sendo convertidos deve ser desligado, sistemas de arquivos devem ser desmontados e grupos de volumes devem ser desativados. Quando o ASM for usado, certifique-se de que o grupo de discos ASM esteja desmontado e desligue todos os serviços de grade.

Desmapear LUNs da rede FC

Depois que os LUNs estiverem totalmente quietos, remova os mapeamentos do grupo FC original.

```
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxhost
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxhost
...
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxhost
Cluster01::*> lun unmap -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxhost
```

Remapear LUNs para a rede IP

Conceda acesso a cada LUN ao novo grupo de iniciadores baseado em iSCSI.

```
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN0 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_asm/LUN1 -igroup
linuxiscsi
...
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN8 -igroup
linuxiscsi
Cluster01::*> lun map -vserver vserver1 -path /vol/new_lvm/LUN9 -igroup
linuxiscsi
Cluster01::*>
```

Descubra iSCSI Targets

Existem duas fases para a descoberta iSCSI. O primeiro é descobrir os alvos, o que não é o mesmo que descobrir um LUN. O `iscsiadm` comando mostrado abaixo sonda o grupo do portal especificado pelo `-p argument` e armazena uma lista de todos os endereços IP e portas que oferecem serviços iSCSI. Neste caso, existem quatro endereços IP que têm serviços iSCSI na porta padrão 3260.

> Este comando pode levar vários minutos para ser concluído se algum dos endereços IP de destino não puder ser alcançado.

```
[root@host1 ~]# iscsiadm -m discovery -t st -p fas8060-iscsi-public1
10.63.147.197:3260,1033 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
10.63.147.198:3260,1034 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
172.20.108.203:3260,1030 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
172.20.108.202:3260,1029 iqn.1992-
08.com.netapp:sn.807615e9ef6111e5a5ae90e2ba5b9464:vs.3
```

Descubra iSCSI LUNs

Depois que os iSCSI Targets forem descobertos, reinicie o serviço iSCSI para descobrir os iSCSI LUNs disponíveis e criar dispositivos associados, como os dispositivos multipath ou ASMlib.

```
[root@host1 ~]# service iscsi restart
Redirecting to /bin/systemctl restart  iscsi.service
```

Reinicie o ambiente

Reinicie o ambiente reativando grupos de volumes, remontando sistemas de arquivos, reiniciando serviços RAC e assim por diante. Como precaução, o NetApp recomenda que você reinicie o servidor após o processo de conversão estar concluído para ter certeza de que todos os arquivos de configuração estão corretos e todos os dispositivos obsoletos são removidos.

Cuidado: Antes de reiniciar um host, certifique-se de que todas as entradas `/etc/fstab` nessa referência migradas dos recursos SAN sejam comentadas. Se esta etapa não for tomada e houver problemas com o acesso LUN, o resultado pode ser um sistema operacional que não inicializa. Este problema não danifica os dados. No entanto, pode ser muito inconveniente inicializar no modo de recuperação ou em um modo semelhante e corrigir `/etc/fstab` para que o sistema operacional possa ser inicializado para permitir que os esforços de solução de problemas comecem.

Exemplos de scripts

Os scripts apresentados são fornecidos como exemplos de como fazer scripts de várias tarefas do sistema operacional e do banco de dados. Eles são fornecidos como estão. Se for necessário suporte para um procedimento específico, entre em Contato com a NetApp ou um revendedor da NetApp.

Encerramento da base de dados

O seguinte script Perl toma um único argumento do Oracle SID e desliga um banco de dados. Ele pode ser executado como o usuário Oracle ou como root.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
77 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF4
sqlplus / as sysdba << EOF2
shutdown immediate;
EOF2
`;};
print @out;
if ("@out" =~ /ORACLE instance shut down/) {
print "$oraclesid shut down\n";
exit 0;}
elsif ("@out" =~ /Connected to an idle instance/) {
print "$oraclesid already shut down\n";
exit 0;}
else {
print "$oraclesid failed to shut down\n";
exit 1;}
```

Inicialização do banco de dados

O seguinte script Perl toma um único argumento do Oracle SID e desliga um banco de dados. Ele pode ser executado como o usuário Oracle ou como root.

```
#! /usr/bin/perl
use strict;
use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
my $uid=$<;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
'
`;}
else {
@out=`. oraenv << EOF3
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
startup;
EOF2
`;};
print @out;
if ("@out" =~ /Database opened/) {
print "$oraclesid started\n";
exit 0;}
elsif ("@out" =~ /cannot start already-running ORACLE/) {
print "$oraclesid already started\n";
exit 1;}
else {
78 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
print "$oraclesid failed to start\n";
exit 1;}
```

Converter sistema de arquivos para somente leitura

O script a seguir toma um argumento file-system e tenta desmontá-lo e remontá-lo como somente leitura. Isso é útil durante os processos de migração nos quais um sistema de arquivos deve ser mantido disponível para replicar dados e, no entanto, deve ser protegido contra danos acidentais.

```
#! /usr/bin/perl
use strict;
#use warnings;
my $filesystem=$ARGV[0];
my @out=`umount '$filesystem'`;
if ($? == 0) {
 print "$filesystem unmounted\n";
 @out = `mount -o ro '$filesystem'`;
 if ($? == 0) {
 print "$filesystem mounted read-only\n";
 exit 0;}}
else {
 print "Unable to unmount $filesystem\n";
 exit 1;}
print @out;
```

Substitua o sistema de arquivos

O exemplo de script a seguir é usado para substituir um sistema de arquivos por outro. Como edita o arquivo "/etc/fstab", ele deve ser executado como root. Ele aceita um único argumento delimitado por vírgulas dos sistemas de arquivos antigos e novos.

1. Para substituir o sistema de arquivos, execute o seguinte script:

```
#! /usr/bin/perl
use strict;
#use warnings;
my $oldfs;
my $newfs;
my @oldfstab;
my @newfstab;
my $source;
my $mountpoint;
my $leftover;
my $oldfstabentry='';
my $newfstabentry='';
my $migratedfstabentry='';
($oldfs, $newfs) = split (',',$ARGV[0]);
open(my $filehandle, '<', '/etc/fstab') or die "Could not open
/etc/fstab\n";
while (my $line = <$filehandle>) {
 chomp $line;
 ($source, $mountpoint, $leftover) = split(/[ , ]/,$line, 3);
 if ($mountpoint eq $oldfs) {
 $oldfstabentry = "#Removed by swap script $source $oldfs $leftover";}
```

```
 elsif ($mountpoint eq $newfs) {
 $newfstabentry = "#Removed by swap script $source $newfs $leftover";
 $migratedfstabentry = "$source $oldfs $leftover";}
 else {
 push (@newfstab, "$line\n")}}
79 Migration of Oracle Databases to NetApp Storage Systems © 2021
NetApp, Inc. All rights reserved
push (@newfstab, "$oldfstabentry\n");
push (@newfstab, "$newfstabentry\n");
push (@newfstab, "$migratedfstabentry\n");
close($filehandle);
if ($oldfstabentry eq ''){
 die "Could not find $oldfs in /etc/fstab\n";}
if ($newfstabentry eq ''){
 die "Could not find $newfs in /etc/fstab\n";}
my @out=`umount '$newfs'`;
if ($? == 0) {
 print "$newfs unmounted\n";}
else {
 print "Unable to unmount $newfs\n";
 exit 1;}
@out=`umount '$oldfs'`;
if ($? == 0) {
 print "$oldfs unmounted\n";}
else {
 print "Unable to unmount $oldfs\n";
 exit 1;}
system("cp /etc/fstab /etc/fstab.bak");
open ($filehandle, ">", '/etc/fstab') or die "Could not open /etc/fstab
for writing\n";
for my $line (@newfstab) {
 print $filehandle $line;}
close($filehandle);
@out=`mount '$oldfs'`;
if ($? == 0) {
 print "Mounted updated $oldfs\n";
 exit 0;}
else{
 print "Unable to mount updated $oldfs\n";
 exit 1;}
exit 0;
```

Como exemplo do uso deste script, suponha que os dados em `/oradata` são migrados para `/neworadata` e `/logs` são migrados para `/newlogs`o . Um dos métodos mais simples para executar esta tarefa é usando uma operação de cópia de arquivo simples para realocar o novo dispositivo de volta para o ponto de montagem original.

2. Suponha que os sistemas de arquivos antigos e novos estão presentes no `/etc/fstab` arquivo da seguinte forma:

```
cluster01:/vol_oradata /oradata nfs rw,bg,vers=3,rsize=65536,wsize=65536
0 0
cluster01:/vol_logs /logs nfs rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_neworadata /neworadata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_newlogs /newlogs nfs rw,bg,vers=3,rsize=65536,wsize=65536
0 0
```

3. Quando executado, esse script desmonta o sistema de arquivos atual e o substitui pelo novo:

```
[root@jfsc3 scripts]# ./swap.fs.pl /oradata,/neworadata
/neworadata unmounted
/oradata unmounted
Mounted updated /oradata
[root@jfsc3 scripts]# ./swap.fs.pl /logs,/newlogs
/newlogs unmounted
/logs unmounted
Mounted updated /logs
```

4. O script também atualiza o `/etc/fstab` arquivo de acordo. No exemplo mostrado aqui, ele inclui as seguintes alterações:

```
#Removed by swap script cluster01:/vol_oradata /oradata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_neworadata /neworadata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_neworadata /oradata nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_logs /logs nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
#Removed by swap script cluster01:/vol_newlogs /newlogs nfs
rw,bg,vers=3,rsize=65536,wsize=65536 0 0
cluster01:/vol_newlogs /logs nfs rw,bg,vers=3,rsize=65536,wsize=65536 0
0
```

Migração automatizada de banco de dados

Este exemplo demonstra o uso de scripts de desligamento, inicialização e substituição do sistema de arquivos para automatizar totalmente uma migração.

```
#! /usr/bin/perl
use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my @oldfs;
my @newfs;
my $x=1;
while ($x < scalar(@ARGV)) {
   ($oldfs[$x-1], $newfs[$x-1]) = split (',',$ARGV[$x]);
   $x+=1;}
my @out=`./dbshut.pl '$oraclesid'`;
print @out;
if ($? ne 0) {
   print "Failed to shut down database\n";
   exit 0;}
$x=0;
while ($x < scalar(@oldfs)) {
   my @out=`./mk.fs.readonly.pl '$oldfs[$x]'`;
   if ($? ne 0) {
      print "Failed to make filesystem $oldfs[$x] readonly\n";
      exit 0;}
   $x+=1;}
$x=0;
while ($x < scalar(@oldfs)) {
   my @out=`rsync -rlpogt --stats --progress --exclude='.snapshot'
'$oldfs[$x]/' '/$newfs[$x]/'`;
   print @out;
   if ($? ne 0) {
      print "Failed to copy filesystem $oldfs[$x] to $newfs[$x]\n";
      exit 0;}
   else {
      print "Succesfully replicated filesystem $oldfs[$x] to
$newfs[$x]\n";}
   $x+=1;}
$x=0;
while ($x < scalar(@oldfs)) {
   print "swap $x $oldfs[$x] $newfs[$x]\n";
   my @out=`./swap.fs.pl '$oldfs[$x],$newfs[$x]'`;
   print @out;
   if ($? ne 0) {
      print "Failed to swap filesystem $oldfs[$x] for $newfs[$x]\n";
      exit 1;}
   else {
      print "Swapped filesystem $oldfs[$x] for $newfs[$x]\n";}
   $x+=1;}
my @out=`./dbstart.pl '$oraclesid'`;
```

```
print @out;
```

Apresentar localizações dos ficheiros

Este script coleta uma série de parâmetros críticos do banco de dados e os imprime em um formato fácil de ler. Este script pode ser útil ao revisar layouts de dados. Além disso, o script pode ser modificado para outros usos.

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;};
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print "$oraclesid datafiles:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
```

```
@out=dosql('select member from v\\\\\$logfile;');
print "$oraclesid redo logs:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select name from v\\\\\$tempfile;');
print "$oraclesid temp datafiles:\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('show parameter spfile;');
print "$oraclesid spfile\n";
for $line (@out) {
        chomp($line);
        if (length($line)>0) {print "$line\n";}}
print "\n";
@out=dosql('select name||\' \'||value from v\\\\\$parameter where
isdefault=\'FALSE\';');
print "$oraclesid key parameters\n";
for $line (@out) {
        chomp($line);
        if ($line =~ /control_files/) {print "$line\n";}
        if ($line =~ /db_create/) {print "$line\n";}
        if ($line =~ /db_file_name_convert/) {print "$line\n";}
        if ($line =~ /log_archive_dest/) {print "$line\n";}}
        if ($line =~ /log_file_name_convert/) {print "$line\n";}
        if ($line =~ /pdb_file_name_convert/) {print "$line\n";}
        if ($line =~ /spfile/) {print "$line\n";}
print "\n";
```

Limpeza da migração do ASM

```
#! /usr/bin/perl
#use strict;
#use warnings;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my @out;
sub dosql{
        my $command = @_[0];
        my @lines;
        my $uid=$<;
        if ($uid == 0) {
```

```
        @lines=`su - $oracleuser -c "export ORAENV_ASK=NO;export
ORACLE_SID=$oraclesid;. oraenv -s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
"
        `;}
        else {
        $command=~s/\\\\\\\/\\/g;
        @lines=`export ORAENV_ASK=NO;export ORACLE_SID=$oraclesid;. oraenv
-s << EOF1
EOF1
sqlplus -S / as sysdba << EOF2
set heading off
$command
EOF2
        `;}
return @lines}
print "\n";
@out=dosql('select name from v\\\\\$datafile;');
print @out;
print "shutdown immediate;\n";
print "startup mount;\n";
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "host mv $line $1$newname\n";}}
print "\n";
for $line (@out) {
        if (length($line) > 1) {
                chomp($line);
                ($first, $second,$third,$fourth)=split('_',$line);
                $fourth =~ s/^TS-//;
                $newname=lc("$fourth.dbf");
                $path2file=$line;
                $path2file=~ /(^.*.\/)/;
                print "alter database rename file '$line' to
'$1$newname';\n";}}
```

```
print "alter database open;\n";
print "\n";
```

ASM para conversão de nome de sistema de arquivos

```
set serveroutput on;
set wrap off;
declare
    cursor df is select file#, name from v$datafile;
    cursor tf is select file#, name from v$tempfile;
    cursor lf is select member from v$logfile;
    firstline boolean := true;
begin
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('Parameters for log file conversion:');
    dbms_output.put_line(CHR(13));
    dbms_output.put('*.log_file_name_convert = ');
    for lfrec in lf loop
        if (firstline = true) then
            dbms_output.put('''' || lfrec.member || ''', ');
            dbms_output.put('''/NEW_PATH/' ||
regexp_replace(lfrec.member,'^.*/','') || '''');
        else
            dbms_output.put(',''' || lfrec.member || ''', ');
            dbms_output.put('''/NEW_PATH/' ||
regexp_replace(lfrec.member,'^.*/','') || '''');
        end if;
        firstline:=false;
    end loop;
    dbms_output.put_line(CHR(13));
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('rman duplication script:');
    dbms_output.put_line(CHR(13));
    dbms_output.put_line('run');
    dbms_output.put_line('{');
    for dfrec in df loop
        dbms_output.put_line('set newname for datafile ' ||
            dfrec.file# || ' to ''' || dfrec.name ||''';');
    end loop;
    for tfrec in tf loop
        dbms_output.put_line('set newname for tempfile ' ||
            tfrec.file# || ' to ''' || tfrec.name ||''';');
    end loop;
    dbms_output.put_line('duplicate target database for standby backup
location INSERT_PATH_HERE;');
    dbms_output.put_line('}');
end;
/
```

Repetir registos na base de dados

Este script aceita um único argumento de um SID Oracle para um banco de dados que está no modo de montagem e tenta reproduzir todos os logs de arquivo disponíveis atualmente.

```
#! /usr/bin/perl
use strict;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
84 Migration of Oracle Databases to NetApp Storage Systems © 2021 NetApp,
Inc. All rights reserved
my $uid = $<;
my @out;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover database until cancel;
auto
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover database until cancel;
auto
EOF2
`;
}
print @out;
```

Repetir registos na base de dados em espera

Esse script é idêntico ao script anterior, exceto que ele é projetado para um banco de dados em espera.

```
#! /usr/bin/perl
use strict;
my $oraclesid=$ARGV[0];
my $oracleuser='oracle';
my $uid = $<;
my @out;
if ($uid == 0) {
@out=`su - $oracleuser -c '. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover standby database until cancel;
auto
EOF2
'
`;}
else {
@out=`. oraenv << EOF1
$oraclesid
EOF1
sqlplus / as sysdba << EOF2
recover standby database until cancel;
auto
EOF2
`;
}
print @out;
```

Notas adicionais

Otimização de desempenho e benchmarking

O teste preciso do desempenho de armazenamento de banco de dados é um assunto extremamente complicado. Requer uma compreensão dos seguintes problemas:

- IOPS e taxa de transferência
- A diferença entre as operações de e/S de primeiro plano e segundo plano
- O efeito da latência sobre o banco de dados
- Várias configurações de sistema operacional e rede que também afetam o desempenho do armazenamento

Além disso, há tarefas de bancos de dados que não são de storage a serem consideradas. Há um ponto em que a otimização do desempenho de storage não produz benefícios úteis porque a performance do storage não é mais um fator limitante para o desempenho.

A maioria dos clientes de bancos de dados agora seleciona all-flash arrays, o que cria algumas considerações adicionais. Por exemplo, considere o teste de desempenho em um sistema AFF A900 de dois nós:

- Com uma taxa de leitura/gravação de 80/20, dois nós de A900 TB podem fornecer mais de 1M IOPS de banco de dados aleatórios antes mesmo de a latência ultrapassar a marca de 150µs. Isso está muito além das demandas atuais de desempenho da maioria dos bancos de dados que é difícil prever a melhoria esperada. O storage seria, em grande parte, apagado como um gargalo.
- A largura de banda da rede é uma fonte cada vez mais comum de limitações de desempenho. Por exemplo, as soluções de disco giratório costumam ser gargalos para a performance do banco de dados, pois a latência de e/S é muito alta. Quando as limitações de latência são removidas por um array all-flash, a barreira frequentemente muda para a rede. Isso é especialmente notável com ambientes virtualizados e sistemas blade, onde a verdadeira conetividade de rede é difícil de visualizar. Isso pode complicar o teste de desempenho se o próprio sistema de armazenamento não puder ser totalmente utilizado devido a limitações de largura de banda.
- Em geral, não é possível comparar o desempenho de um array all-flash com um array que contém discos giratórios devido à latência drasticamente aprimorada dos all-flash arrays. Os resultados dos testes normalmente não são significativos.
- Comparar o desempenho máximo de IOPS com um array all-flash geralmente não é um teste útil, pois os bancos de dados não são limitados pela e/S do storage Por exemplo, suponha que um array possa suportar 500K IOPS aleatórios, enquanto outro pode sustentar 300KK. A diferença é irrelevante no mundo real se um banco de dados está gastando 99% do seu tempo em processamento de CPU. Os workloads nunca utilizam todas as funcionalidades do storage array. Em contraste, os recursos de IOPS de pico podem ser críticos em uma plataforma de consolidação na qual o storage array deve ser carregado para seus recursos de pico.
- Considere sempre a latência e o IOPS em qualquer teste de storage. Muitos storage arrays no mercado reivindicam níveis extremos de IOPS, mas a latência torna esses IOPS inúteis em tais níveis. O destino típico com all-flash arrays é a marca 1ms. Uma abordagem melhor para testar não é medir o máximo de IOPS possível, mas determinar quantos IOPS um storage array pode sustentar antes que a latência média seja maior que 1ms.

Oracle Automatic Workload Repository e benchmarking

O padrão-ouro para comparações de desempenho Oracle é um relatório do Oracle Automatic Workload Repository (AWR).

Existem vários tipos de relatórios AWR. Do ponto de vista de armazenamento, um relatório gerado pela execução do `awrrpt.sql` comando é o mais abrangente e valioso, porque tem como alvo uma instância de banco de dados específica e inclui alguns histogramas detalhados que quebram eventos de e/S de armazenamento com base na latência.

Comparar dois arrays de desempenho idealmente envolve executar a mesma carga de trabalho em cada array e produzir um relatório AWR que segmente precisamente a carga de trabalho. No caso de uma carga de trabalho muito longa, um único relatório de AWR com um tempo decorrido que engloba o tempo de início e de parada pode ser usado, mas é preferível dividir os dados de AWR como vários relatórios. Por exemplo, se um trabalho em lote foi executado da meia-noite às 6 da manhã, crie uma série de relatórios AWR de uma hora das meia-noite às 1 da manhã, das 1 às 2 da manhã, e assim por diante.

Em outros casos, uma consulta muito curta deve ser otimizada. A melhor opção é um relatório AWR baseado em um instantâneo AWR criado quando a consulta começa e um segundo instantâneo AWR criado quando a consulta termina. O servidor de banco de dados deve ser silencioso para minimizar a atividade em segundo plano que obscureceria a atividade da consulta em análise.

Onde os relatórios AWR não estão disponíveis, os relatórios do Oracle statspack são uma boa alternativa. Eles contêm a maioria das mesmas estatísticas de e/S que um relatório AWR.

Oracle AWR e solução de problemas

Um relatório AWR também é a ferramenta mais importante para analisar um problema de desempenho.

Assim como no benchmarking, a solução de problemas de desempenho exige que você meça com precisão uma carga de trabalho específica. Quando possível, forneça dados de AWR ao relatar um problema de desempenho ao centro de suporte da NetApp ou ao trabalhar com um NetApp ou equipe de conta de parceiro sobre uma nova solução.

Ao fornecer dados AWR, considere os seguintes requisitos:

- Execute o `awrrpt.sql` comando para gerar o relatório. A saída pode ser texto ou HTML.
- Se os Oracle Real Application clusters (RACs) forem usados, gere relatórios AWR para cada instância no cluster.
- Segmente a hora específica em que o problema existia. O tempo decorrido máximo aceitável de um relatório AWR é geralmente de uma hora. Se um problema persistir por várias horas ou envolver uma operação de várias horas, como um trabalho em lote, forneça vários relatórios AWR de uma hora que cobrem todo o período a ser analisado.
- Se possível, ajuste o intervalo de instantâneos AWR para 15 minutos. Esta definição permite efetuar uma análise mais detalhada. Isso também requer execuções adicionais de `awrrpt.sql` para fornecer um relatório para cada intervalo de 15 minutos.
- Se o problema for uma consulta em execução muito curta, forneça um relatório AWR com base em um instantâneo AWR criado quando a operação começar e um segundo instantâneo AWR criado quando a operação terminar. O servidor de banco de dados deve ser silencioso para minimizar a atividade em segundo plano que obscureceria a atividade da operação em análise.
- Se um problema de desempenho for relatado em determinados momentos, mas não em outros, forneça dados AWR adicionais que demonstrem bom desempenho para comparação.

calibrar_io

O `calibrate_io` comando nunca deve ser usado para testar, comparar ou comparar sistemas de storage. Conforme indicado na documentação da Oracle, este procedimento calibra os recursos de e/S do armazenamento.

A calibração não é a mesma que o benchmarking. O objetivo deste comando é emitir e/S para ajudar a calibrar as operações do banco de dados e melhorar sua eficiência otimizando o nível de e/S emitido para o host. Como o tipo de I/o realizado pela `calibrate_io` operação não representa a e/S real do usuário do banco de dados, os resultados não são previsíveis e frequentemente nem reprodutíveis.

SLOB2

SLOB2, o silly Little Oracle Benchmark, tornou-se a ferramenta preferida para avaliar o desempenho do banco de dados. Foi desenvolvido por Kevin Closson e está disponível em "https://kevinclosson.net/slob/". Leva minutos para instalar e configurar, e ele usa um banco de dados Oracle real para gerar padrões de e/S em um espaço de tabela definido pelo usuário. É uma das poucas opções de teste disponíveis que pode saturar um array all-flash com e/S. Também é útil para gerar níveis muito mais baixos de e/S para simular workloads de storage com IOPS baixo, mas sensíveis à latência.

Banco oscilante

O Swingbench pode ser útil para testar o desempenho do banco de dados, mas é extremamente difícil usar o Swingbench de uma forma que estressa o armazenamento. NetApp não viu nenhum teste do Swingbench que rendeu e/S suficiente para ser uma carga significativa em qualquer array AFF. Em casos limitados, o teste de entrada de pedidos (OET) pode ser usado para avaliar o armazenamento de um ponto de vista de latência. Isso pode ser útil em situações em que um banco de dados tem uma dependência de latência conhecida para consultas específicas. É preciso ter cuidado para garantir que o host e a rede estejam configurados adequadamente para realizar os potenciais de latência de um array all-flash.

HammerDB

HammerDB é uma ferramenta de teste de banco de dados que simula benchmarks TPC-C e TPC-H, entre outros. Pode levar muito tempo para construir um conjunto de dados suficientemente grande para executar corretamente um teste, mas pode ser uma ferramenta eficaz para avaliar o desempenho para aplicativos OLTP e data warehouse.

Orion

A ferramenta Oracle Orion foi comumente usada com o Oracle 9, mas não foi mantida para garantir a compatibilidade com alterações em vários sistemas operacionais de host. Raramente é usado com Oracle 10i ou Oracle 11i devido a incompatibilidades com o SO e configuração de armazenamento.

A Oracle reescreveu a ferramenta, e ela é instalada por padrão com o Oracle 12c. Embora este produto tenha sido melhorado e use muitas das mesmas chamadas que um banco de dados Oracle real usa, ele não usa exatamente o mesmo caminho de código ou comportamento de e/S usado pela Oracle. Por exemplo, a maioria das e/S Oracle são executadas de forma síncrona, o que significa que o banco de dados pára até que a e/S esteja concluída à medida que a operação de e/S for concluída em primeiro plano. Simplesmente inundar um sistema de armazenamento com e/S aleatórias não é uma reprodução de e/S Oracle real e não oferece um método direto de comparar matrizes de armazenamento ou medir o efeito das alterações de configuração.

Dito isto, existem alguns casos de uso para Orion, como a medição geral do desempenho máximo possível de uma configuração particular de armazenamento de rede-host, ou para medir a integridade de um sistema de armazenamento. Com testes cuidadosos, testes Orion utilizáveis podem ser desenvolvidos para comparar matrizes de armazenamento ou avaliar o efeito de uma alteração de configuração, desde que os parâmetros incluam consideração de IOPS, taxa de transferência e latência e tentativa de replicar fielmente uma carga de trabalho realista.

Stale NFSv3 fechaduras

Se um servidor de banco de dados Oracle falhar, ele pode ter problemas com bloqueios NFS obsoletos ao reiniciar. Este problema é evitável prestando atenção cuidadosa à configuração da resolução de nomes no servidor.

Este problema surge porque criar um bloqueio e limpar um bloqueio usam dois métodos ligeiramente diferentes de resolução de nomes. Dois processos estão envolvidos, o Network Lock Manager (NLM) e o cliente NFS. O NLM usa `uname n` para determinar o nome do host, enquanto o `rpc.statd` processo usa `gethostbyname()`o . Esses nomes de host devem corresponder para que o sistema operacional limpe corretamente os bloqueios obsoletos. Por exemplo, o host pode estar procurando bloqueios de propriedade `dbserver5` do , mas os bloqueios foram registrados pelo host como `dbserver5.mydomain.org`. Se `gethostbyname()` não retornar o mesmo valor que `uname -a`, o processo de liberação do bloqueio não foi bem-sucedido.

O script de exemplo a seguir verifica se a resolução do nome é totalmente consistente:

```
#! /usr/bin/perl
$uname=`uname -n`;
chomp($uname);
($name, $aliases, $addrtype, $length, @addrs) = gethostbyname $uname;
print "uname -n yields: $uname\n";
print "gethostbyname yields: $name\n";
```

Se `gethostbyname` não corresponder `uname`, é provável que os bloqueios obsoletos. Por exemplo, este resultado revela um problema potencial:

```
uname -n yields: dbserver5
gethostbyname yields: dbserver5.mydomain.org
```

A solução é geralmente encontrada alterando a ordem em que os hosts aparecem em `/etc/hosts`. por exemplo, suponha que o arquivo hosts inclua esta entrada:

```
10.156.110.201  dbserver5.mydomain.org dbserver5 loghost
```

Para resolver esse problema, altere a ordem em que o nome de domínio totalmente qualificado e o nome de host curto aparecem:

```
10.156.110.201  dbserver5 dbserver5.mydomain.org loghost
```

`gethostbyname()` agora retorna o nome de host curto `dbserver5`, que corresponde à saída `uname` do . Assim, os bloqueios são apagados automaticamente após uma falha do servidor.

Verificação do alinhamento do WAFL

O alinhamento correto do WAFL é essencial para um bom desempenho. Embora o ONTAP gere blocos em 4KB unidades, esse fato não significa que o ONTAP realize todas as operações em 4KB unidades. Na verdade, o ONTAP suporta operações de blocos de diferentes tamanhos, mas a contabilidade subjacente é gerenciada pela WAFL em 4KB unidades.

O termo "alinhamento" refere-se a como Oracle I/o corresponde a essas 4KB unidades. O desempenho ideal requer um bloco Oracle 8KBi para residir em dois blocos físicos de 4KB WAFL em uma unidade. Se um bloco é compensado por 2KB, este bloco reside em metade de um bloco 4KB, um bloco 4KB completo separado e, em seguida, metade de um terceiro bloco 4KB. Este arranjo causa degradação do desempenho.

O alinhamento não é um problema com sistemas de arquivos nas. Os arquivos de dados Oracle estão alinhados ao início do arquivo com base no tamanho do bloco Oracle. Portanto, os tamanhos de bloco de 8KB, 16KB e 32KB estão sempre alinhados. Todas as operações de bloco são compensadas desde o início do arquivo em unidades de 4 kilobytes.

Os LUNs, em contraste, geralmente contêm algum tipo de cabeçalho de driver ou metadados do sistema de arquivos no início que cria um deslocamento. O alinhamento raramente é um problema em sistemas operacionais modernos porque esses sistemas operacionais são projetados para unidades físicas que podem usar um setor 4KB nativo, que também requer que e/S sejam alinhados aos limites 4KB para um desempenho ideal.

Há, no entanto, algumas exceções. Um banco de dados pode ter sido migrado de um sistema operacional antigo que não foi otimizado para e/S 4KB, ou erro de usuário durante a criação da partição pode ter levado a um deslocamento que não está em unidades de tamanho 4KB.

Os exemplos a seguir são específicos do Linux, mas o procedimento pode ser adaptado para qualquer sistema operacional.

Alinhado

O exemplo a seguir mostra uma verificação de alinhamento em um único LUN com uma única partição.

Primeiro, crie a partição que usa todas as partições disponíveis na unidade.

```
[root@host0 iscsi]# fdisk /dev/sdb
Device contains neither a valid DOS partition table, nor Sun, SGI or OSF
disklabel
Building a new DOS disklabel with disk identifier 0xb97f94c1.
Changes will remain in memory only, until you decide to write them.
After that, of course, the previous content won't be recoverable.
The device presents a logical sector size that is smaller than
the physical sector size. Aligning to a physical sector (or optimal
I/O) size boundary is recommended, or performance may be impacted.
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-10240, default 1):
Using default value 1
Last cylinder, +cylinders or +size{K,M,G} (1-10240, default 10240):
Using default value 10240
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
[root@host0 iscsi]#
```

O alinhamento pode ser verificado matematicamente com o seguinte comando:

```
[root@host0 iscsi]# fdisk -u -l /dev/sdb
Disk /dev/sdb: 10.7 GB, 10737418240 bytes
64 heads, 32 sectors/track, 10240 cylinders, total 20971520 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 4096 bytes
I/O size (minimum/optimal): 4096 bytes / 65536 bytes
Disk identifier: 0xb97f94c1
   Device Boot      Start         End      Blocks   Id  System
/dev/sdb1              32    20971519    10485744   83  Linux
```

A saída mostra que as unidades são 512 bytes e o início da partição é 32 unidades. Este é um total de 32 x 512 de 16.834 bytes, que é um múltiplo inteiro de 4KB blocos WAFL. Esta partição está corretamente alinhada.

Para verificar o alinhamento correto, execute as seguintes etapas:

1. Identifique o identificador universal único (UUID) do LUN.

```
FAS8040SAP::> lun show -v /vol/jfs_luns/lun0
              Vserver Name: jfs
                  LUN UUID: ed95d953-1560-4f74-9006-85b352f58fcd
                    Mapped: mapped`                `
```

2. Insira o shell do nó no controlador ONTAP.

```
FAS8040SAP::> node run -node FAS8040SAP-02
Type 'exit' or 'Ctrl-D' to return to the CLI
FAS8040SAP-02> set advanced
set not found.  Type '?' for a list of commands
FAS8040SAP-02> priv set advanced
Warning: These advanced commands are potentially dangerous; use
         them only when directed to do so by NetApp
         personnel.
```

3. Inicie coleções estatísticas no UUID alvo identificado no primeiro passo.

```
FAS8040SAP-02*> stats start lun:ed95d953-1560-4f74-9006-85b352f58fcd
Stats identifier name is 'Ind0xffffff08b9536188'
FAS8040SAP-02*>
```

4. Execute algumas I/O.. É importante usar o `iflag` argumento para garantir que a e/S seja síncrona e não armazenada em buffer.

Tenha muito cuidado com este comando. Reverter os `if` argumentos e `of` destrói os dados.

```
[root@host0 iscsi]# dd if=/dev/sdb1 of=/dev/null iflag=dsync count=1000
bs=4096
1000+0 records in
1000+0 records out
4096000 bytes (4.1 MB) copied, 0.0186706 s, 219 MB/s
```

5. Pare as estatísticas e veja o histograma de alinhamento. Toda a e/S deve estar no `.0` balde, o que indica e/S que está alinhada a um limite de bloco 4KB.

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff08b9536188
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-
4f74-9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.0:186%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.1:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.2:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.3:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.4:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.5:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.6:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.7:0%
```

Desalinhado

O exemplo a seguir mostra e/S desalinhadas:

1. Crie uma partição que não esteja alinhada a um limite 4KB. Este não é um comportamento padrão em sistemas operacionais modernos.

```
[root@host0 iscsi]# fdisk -u /dev/sdb
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (32-20971519, default 32): 33
Last sector, +sectors or +size{K,M,G} (33-20971519, default 20971519):
Using default value 20971519
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
```

2. A partição foi criada com um deslocamento de 33 setores em vez do padrão 32. Repita o procedimento descrito em "Alinhado". O histograma aparece da seguinte forma:

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff0468242e78
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-
4f74-9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.0:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.1:136%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.2:4%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.3:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.4:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.5:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.6:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_align_histo.7:0%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:read_partial_blocks:31%
```

O desalinhamento é claro. A e/S cai principalmente no bucket**`.1`, que corresponde ao deslocamento esperado. Quando a partição foi criada, ela foi movida 512 bytes mais para o dispositivo do que o padrão otimizado, o que significa que o histograma é deslocado em 512 bytes.

Além disso, a `read_partial_blocks` estatística é diferente de zero, o que significa que I/o foi realizado que não preencheu um bloco 4KB inteiro.

Refazer o registo

Os procedimentos aqui explicados são aplicáveis aos datafiles. Os logs do Oracle refazer e os logs de arquivamento têm padrões de e/S diferentes. Por exemplo, refazer o Registro é uma substituição circular de um único arquivo. Se o tamanho padrão de bloco de 512 bytes for usado, as estatísticas de gravação são semelhantes a isso:

```
FAS8040SAP-02*> stats stop
StatisticsID: Ind0xffffff0468242e78
lun:ed95d953-1560-4f74-9006-85b352f58fcd:instance_uuid:ed95d953-1560-4f74-
9006-85b352f58fcd
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.0:12%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.1:8%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.2:4%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.3:10%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.4:13%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.5:6%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.6:8%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_align_histo.7:10%
lun:ed95d953-1560-4f74-9006-85b352f58fcd:write_partial_blocks:85%
```

A e/S seria distribuída em todos os intervalos de histograma, mas isso não é uma preocupação de desempenho. No entanto, as taxas de refazer o log extremamente altas podem se beneficiar do uso de um tamanho de bloco de 4KBMB. Nesse caso, é desejável garantir que os LUNs de refazer o log estejam alinhados corretamente. No entanto, isso não é tão crítico para um bom desempenho como o alinhamento de arquivos de dados.

Informações sobre direitos autorais

Informações sobre marcas comerciais