MetroCluster

Enterprise applications

NetApp
May 09, 2024

목차

MetroCluster . . . 1
MetroCluster 물리적 아키텍처 및 Oracle 데이터베이스 . . . 1
MetroCluster 논리적 아키텍처와 Oracle 데이터베이스 . . . 5
Oracle 데이터베이스 및 SyncMirror . . . 11
MetroCluster을 사용한 Oracle 데이터베이스 페일오버 . . . 12
모든 주요 MetroCluster 파이버 채널 인프라 환경을 지원합니다 . . . 13
MetroCluster에 있는 Oracle 단일 인스턴스 . . . 15
MetroCluster에서 Oracle RAC 확장 . . . 16

MetroCluster

MetroCluster 물리적 아키텍처 및 Oracle 데이터베이스

MetroCluster 환경에서 Oracle 데이터베이스가 작동하는 방식을 이해하려면 MetroCluster 시스템의 물리적 설계에 대해 몇 가지 설명이 필요합니다.

이 문서는 이전에 게시된 기술 보고서_TR-4592: MetroCluster 기반 Oracle._ 을(를) 대체합니다

MetroCluster는 3가지 구성으로 사용할 수 있습니다

- IP 연결이 포함된 HA 쌍
- FC 연결이 포함된 HA 쌍
- FC 연결이 포함된 단일 컨트롤러

[참고] '접속'이라는 용어는 사이트 간 복제에 사용되는 클러스터 접속을 의미합니다. 호스트 프로토콜을 참조하지 않습니다. 모든 호스트측 프로토콜은 클러스터 간 통신에 사용되는 연결 유형에 관계없이 MetroCluster 구성에서 평소와 같이 지원됩니다.

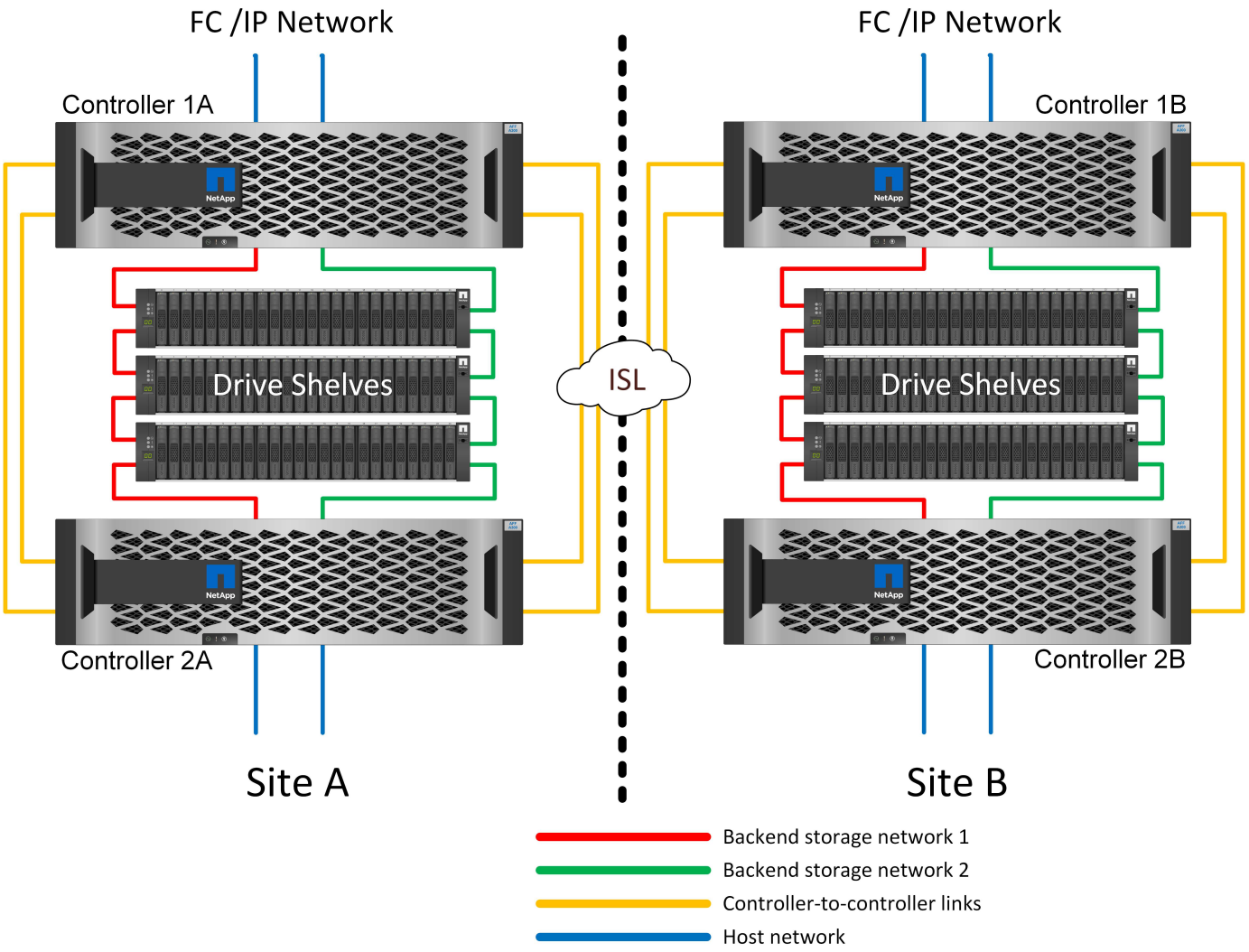
MetroCluster IP를 선택합니다

HA 쌍 MetroCluster IP 구성은 사이트당 2~4개의 노드를 사용합니다. 이 구성 옵션은 2노드 옵션에 비해 복잡성과 비용을 증가시키지만, 내부 중복이라는 중요한 이점을 제공합니다. 컨트롤러 장애가 간단하더라도 WAN을 통한 데이터 액세스가 필요하지 않습니다. 데이터 액세스는 대체 로컬 컨트롤러를 통해 로컬에 유지됩니다.

대부분의 고객은 인프라 요구 사항이 더 간단하기 때문에 IP 연결을 선택하고 있습니다. 과거에는 다크 파이버 및 FC 스위치를 사용하여 고속 사이트 간 연결을 제공하기가 일반적으로 더 쉬웠지만, 오늘날의 고속, 짧은 지연 시간 IP 회로는 보다 쉽게 사용할 수 있었습니다.

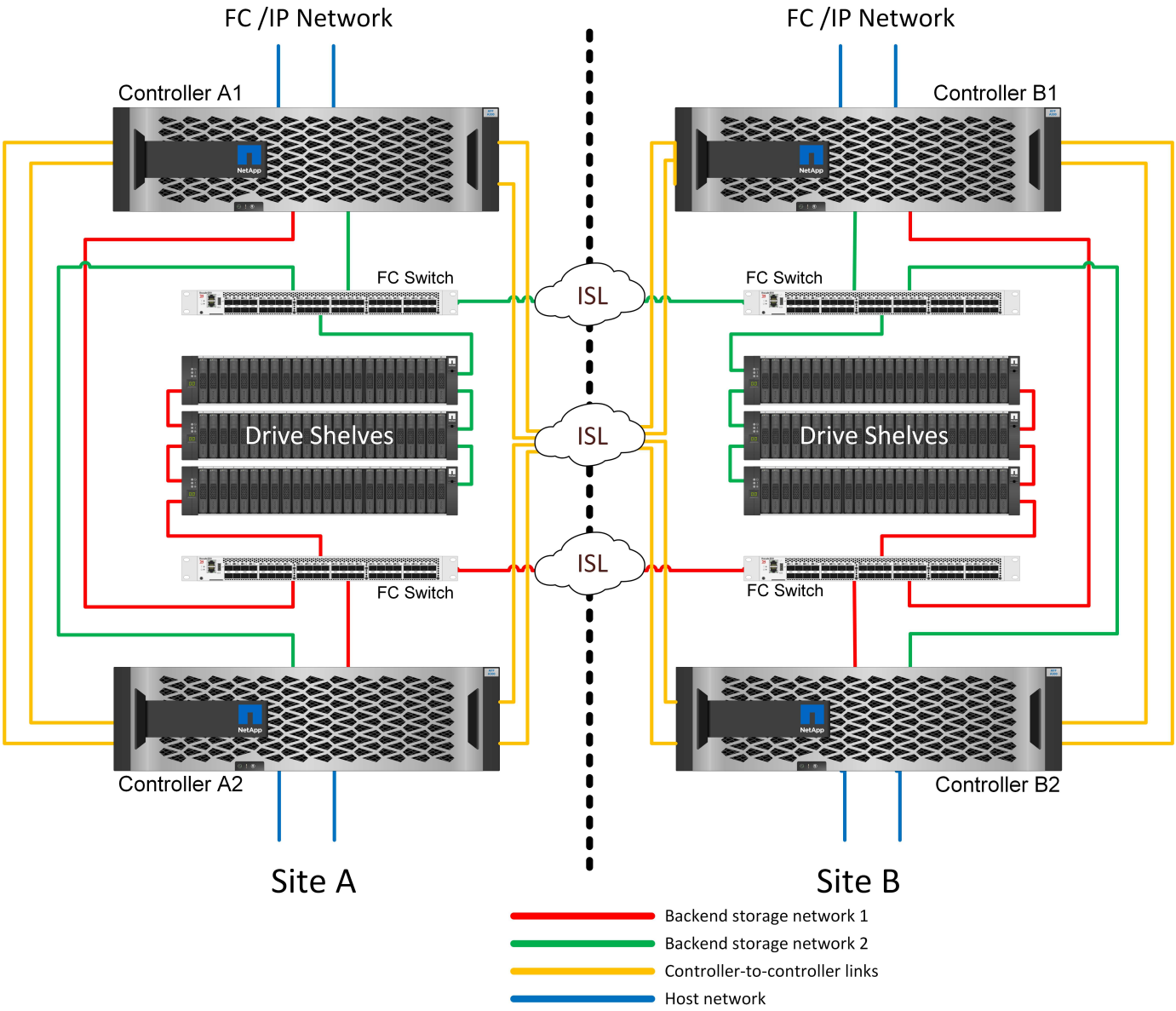
또한 사이트 간 연결만 컨트롤러를 위한 것이므로 아키텍처가 더욱 단순합니다. FC SAN 연결 MetroCluster에서 컨트롤러는 반대쪽 사이트의 드라이브에 직접 기록하므로 SAN 연결, 스위치 및 브리지가 추가로 필요합니다. 반면, IP 구성의 컨트롤러는 컨트롤러를 통해 반대쪽 드라이브에 씁니다.

자세한 내용은 공식 ONTAP 설명서 및 를 참조하십시오 "MetroCluster IP 솔루션 아키텍처 및 설계".



HA-쌍 FC SAN 연결 MetroCluster

HA 쌍 MetroCluster FC 구성은 사이트당 2개 또는 4개의 노드를 사용합니다. 이 구성 옵션은 2노드 옵션에 비해 복잡성과 비용을 증가시키지만, 내부 중복이라는 중요한 이점을 제공합니다. 컨트롤러 장애가 간단하더라도 WAN을 통한 데이터 액세스가 필요하지 않습니다. 데이터 액세스는 대체 로컬 컨트롤러를 통해 로컬에 유지됩니다.

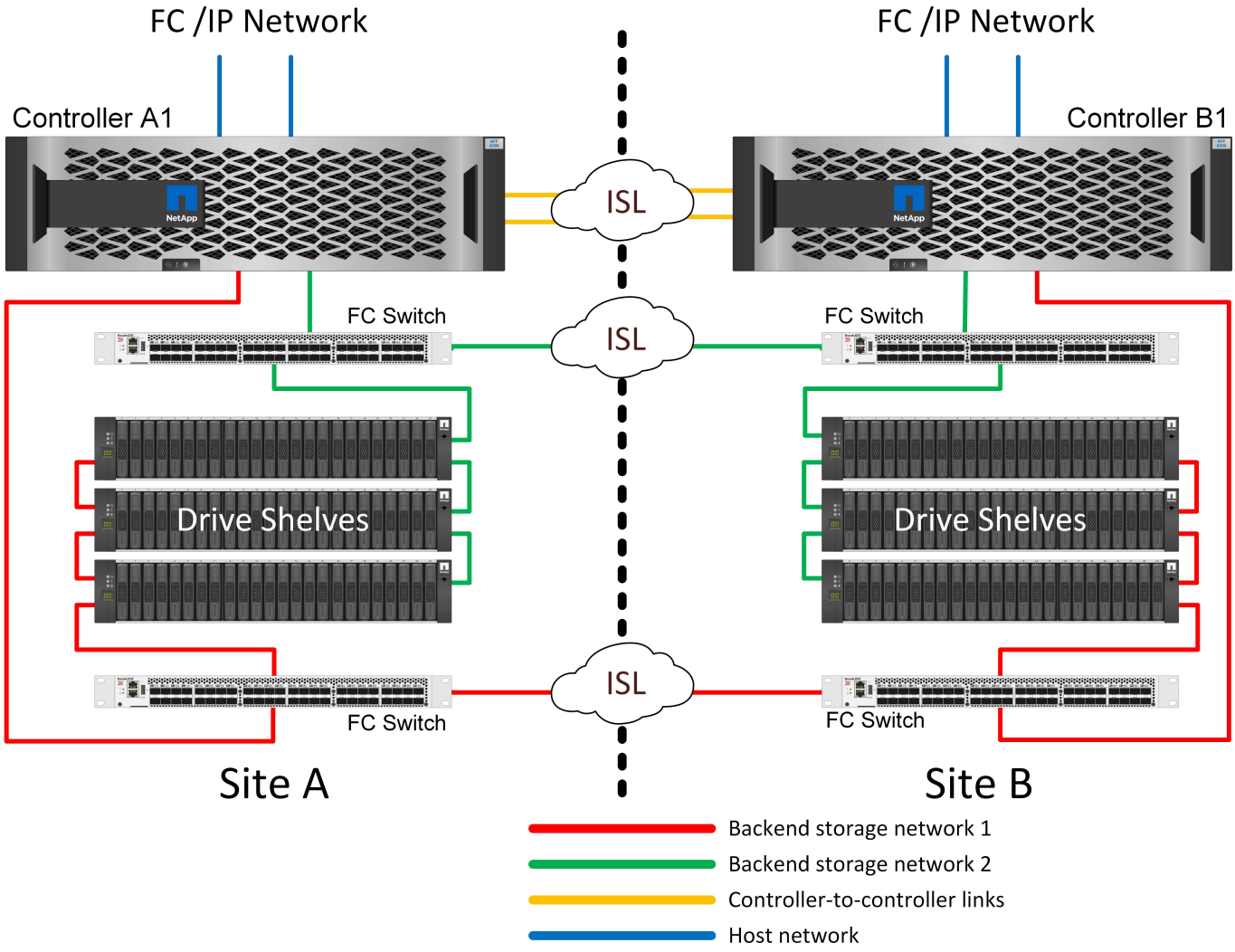


일부 멀티사이트 인프라는 액티브-액티브 운영을 위해 설계되지 않았지만 운영 사이트 및 재해 복구 사이트로 더 많이 사용됩니다. 이 상황에서 HA 쌍 MetroCluster 옵션이 일반적으로 다음과 같은 이유로 더 권장됩니다.

- 2노드 MetroCluster 클러스터는 HA 시스템이지만, 컨트롤러의 예상치 못한 장애나 계획된 유지 관리를 위해서는 반대쪽 사이트에서 데이터 서비스를 온라인으로 전환해야 합니다. 사이트 간 네트워크 연결이 필요한 대역폭을 지원할 수 없는 경우 성능이 영향을 받습니다. 유일한 옵션은 다양한 호스트 OS 및 관련 서비스를 대체 사이트로 페일오버하는 것입니다. HA Pair MetroCluster 클러스터는 동일한 사이트 내에서 단순한 페일오버가 발생하기 때문에 이 문제가 해소됩니다.
- 일부 네트워크 토폴로지는 사이트 간 액세스용으로 설계되지 않은 대신 서로 다른 서브넷이나 격리된 FC SAN을 사용합니다. 이런 경우 2노드 MetroCluster 클러스터는 다른 사이트의 서버에 데이터를 제공할 수 없기 때문에 더 이상 HA 시스템으로 작동하지 않습니다. 완벽한 이중화를 제공하려면 HA Pair MetroCluster 옵션이 필요합니다.
- 2개 사이트 인프라를 고가용성 단일 인프라로 간주하는 경우 2노드 MetroCluster 구성이 적합합니다. 하지만 사이트 장애 후 시스템이 오랫동안 작동해야 하는 경우에는 단일 사이트 내에서 HA를 계속 제공하기 때문에 HA 2노드가 선호됩니다.

2노드 FC SAN 연결 MetroCluster

2노드 MetroCluster 구성은 사이트당 하나의 노드만 사용합니다. 이 설계는 구성과 유지 관리가 필요한 구성 요소가 적기 때문에 HA 쌍 옵션보다 단순합니다. 또한 케이블 연결 및 FC 스위칭에 대한 인프라 요구도 줄었습니다. 마지막으로 비용을 절감할 수 있습니다.



이 설계의 분명한 영향은 단일 사이트에서 컨트롤러 장애가 발생하면 반대쪽 사이트에서 데이터를 사용할 수 있다는 것입니다. 이러한 제한이 반드시 문제가 되는 것은 아닙니다. 많은 기업은 기본적으로 단일 인프라로 작동하는 지연 시간이 짧은 확장된 고속 네트워크를 통해 멀티사이트 데이터 센터를 운영하고 있습니다. 이 경우 MetroCluster의 2노드 버전을 사용하는 것이 좋습니다. 현재 여러 서비스 공급자가 두 노드 시스템을 페타바이트 규모로 사용하고 있습니다.

MetroCluster 복원력 기능

MetroCluster 솔루션에는 단일 장애 지점이 없습니다.

- 각 컨트롤러에는 로컬 사이트의 드라이브 쉘프에 대한 2개의 독립적 경로가 있습니다.
- 각 컨트롤러에는 원격 사이트의 드라이브 쉘프에 대한 두 개의 독립적 경로가 있습니다.
- 각 컨트롤러에는 반대쪽 사이트에 있는 컨트롤러에 대한 독립적인 경로가 2개 있습니다.
- HA 쌍 구성에서 각 컨트롤러에는 로컬 파트너에 대한 두 가지 경로가 있습니다.

요약하면, MetroCluster의 데이터 제공 기능에 영향을 주지 않으면서 구성의 모든 구성 요소를 제거할 수 있습니다. 두 옵션 간의 복원력에서 유일한 차이점은 HA 쌍 버전이 사이트 장애 발생 후 전체 HA 스토리지 시스템이라는 점입니다.

MetroCluster 논리적 아키텍처와 Oracle 데이터베이스

MetroCluster 환경에서 Oracle 데이터베이스가 작동하는 방식을 이해하려면 MetroCluster 시스템의 논리적 기능에 대한 몇 가지 설명이 필요합니다.

사이트 장애 방지: NVRAM 및 MetroCluster

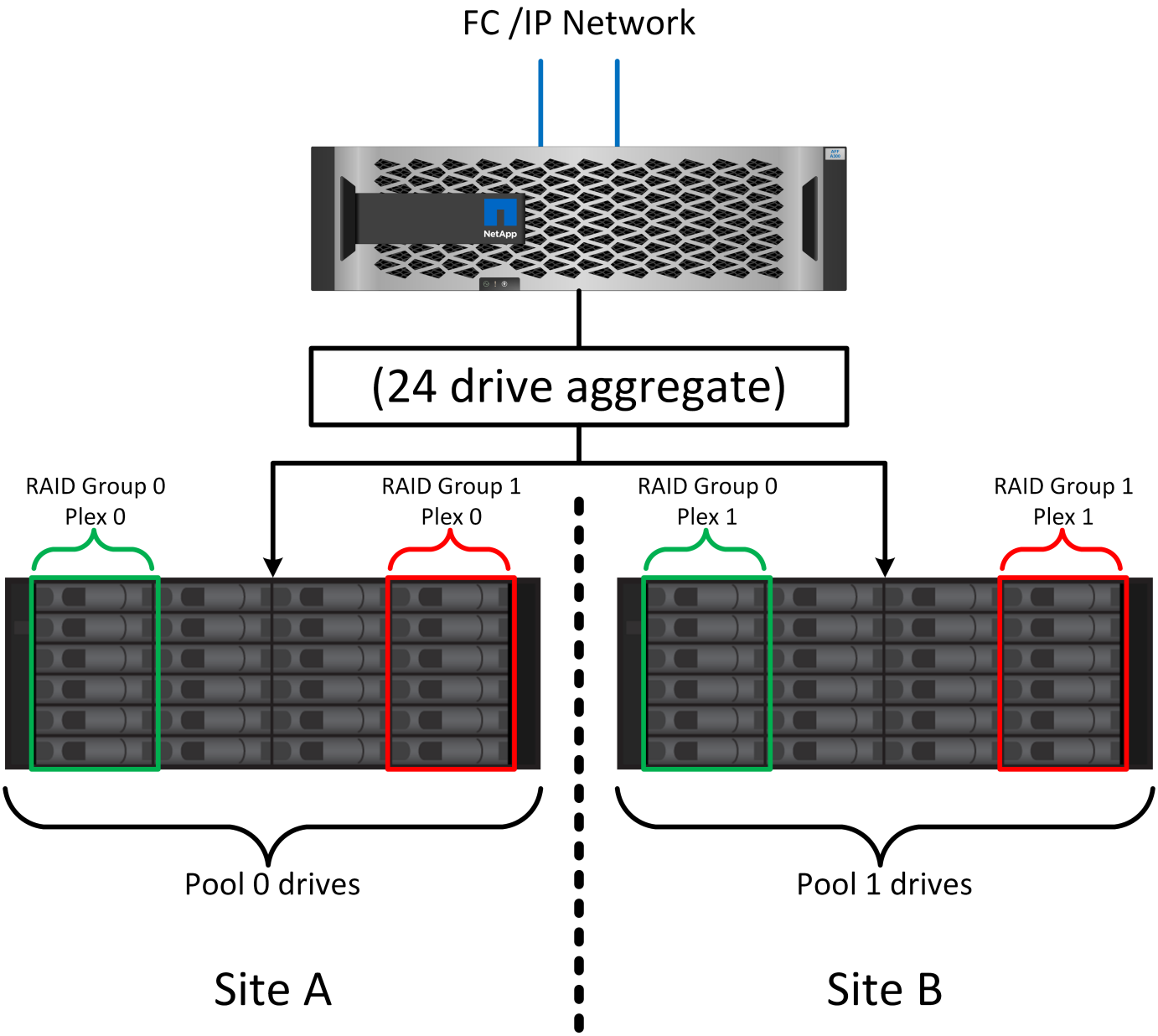
MetroCluster는 NVRAM 데이터 보호를 다음과 같은 방식으로 확장합니다.

- 2노드 구성에서는 ISL(Inter-Switch Link)을 사용하여 NVRAM 데이터를 원격 파트너에게 복제합니다.
- HA 쌍 구성에서는 NVRAM 데이터가 로컬 파트너와 원격 파트너 모두에 복제됩니다.
- 쓰기가 모든 파트너에게 복제될 때까지 확인되지 않습니다. 이 아키텍처는 NVRAM 데이터를 원격 파트너에게 복제하여 전송 중인 I/O를 사이트 장애로부터 보호합니다. 이 프로세스는 드라이브 수준 데이터 복제와 관련되지 않습니다. 애그리게이트를 소유한 컨트롤러는 애그리게이트의 두 플렉스에 쓰기를 통해 데이터 복제를 담당하지만, 사이트 손실 시 전송 중인 I/O 손실에 대한 보호가 여전히 필요합니다. 복제된 NVRAM 데이터는 파트너 컨트롤러가 장애가 발생한 컨트롤러를 인수해야 하는 경우에만 사용됩니다.

사이트 및 쉘프 장애 보호: SyncMirror 및 플렉스

SyncMirror는 RAID DP 또는 RAID-TEC를 향상하지만 대체하지는 않는 미러링 기술입니다. 2개의 독립적인 RAID 그룹의 콘텐츠를 미러링합니다. 논리적 구성은 다음과 같습니다.

1. 드라이브는 위치에 따라 두 개의 풀로 구성됩니다. 하나의 풀은 사이트 A의 모든 드라이브로 구성되고, 두 번째 풀은 사이트 B의 모든 드라이브로 구성됩니다
2. 그런 다음 애그리게이트라고 하는 공통 스토리지 풀이 RAID 그룹의 미러링된 세트를 기반으로 생성됩니다. 각 사이트에서 동일한 수의 드라이브가 그려집니다. 예를 들어, 20개 드라이브로 구성된 SyncMirror 애그리게이트는 사이트 A의 드라이브 10개와 사이트 B의 드라이브 10개로 구성됩니다
3. 특정 사이트의 각 드라이브 세트는 미러링 사용과 상관없이 하나 이상의 완전히 이중화된 RAID DP 또는 RAID-TEC 그룹으로 자동으로 구성됩니다. 이와 같이 미러링에서 RAID를 사용하면 사이트 손실 후에도 데이터를 보호할 수 있습니다.



위 그림은 SyncMirror 구성의 예를 보여 줍니다. 24-드라이브 애그리게이트가 사이트 A에 할당된 쉘프의 드라이브 12개와 사이트 B에 할당된 쉘프의 드라이브 12개로 컨트롤러에서 생성되었습니다 드라이브는 두 개의 미러링된 RAID 그룹으로 그룹화되었습니다. RAID 그룹 0에는 사이트 A의 6개 드라이브 플렉스가 사이트 B의 6개 드라이브 플렉스에 미러링됩니다 마찬가지로, RAID 그룹 1에는 사이트 A의 6개 드라이브 플렉스가 사이트 B의 6개 드라이브 플렉스에 미러링됩니다

SyncMirror는 일반적으로 각 사이트에 하나의 데이터 복사본으로 MetroCluster 시스템에 원격 미러링을 제공하는 데 사용됩니다. 경우에 따라 단일 시스템에서 추가 수준의 이중화를 제공하기 위해 사용되었습니다. 특히, 쉘프 레벨 이중화를 제공합니다. 드라이브 쉘프에는 이미 이중 전원 공급 장치와 컨트롤러가 포함되어 있으며 전반적으로 판금보다 조금 더 크지만, 경우에 따라 추가 보호가 필요할 수 있습니다. 예를 들어, 한 NetApp 고객은 자동차 테스트에 사용되는 모바일 실시간 분석 플랫폼용 SyncMirror를 구축했습니다. 시스템은 독립적인 전원 공급 장치와 독립적인 UPS 시스템이 함께 제공되는 두 개의 물리적 랙으로 분리되었습니다.

이중화 실패: **NVFAIL**

앞서 설명한 것처럼, 쓰기는 하나 이상의 다른 컨트롤러에서 로컬 NVRAM 및 NVRAM에 로그인되기 전까지는 승인되지

않습니다. 이렇게 하면 하드웨어 장애나 정전이 발생해도 전송 중인 I/O가 손실되지 않습니다 로컬 NVRAM에 장애가 발생하거나 다른 노드에 대한 연결이 실패하면 데이터가 더 이상 미러링되지 않습니다.

로컬 NVRAM에 오류가 보고되면 노드가 종료됩니다. 이 종료를 통해 HA Pair를 사용할 경우 파트너 컨트롤러로 페일오버됩니다. MetroCluster를 사용할 경우 선택한 전체 구성에 따라 동작이 달라지지만 원격 메모로 자동 페일오버될 수 있습니다. 오류가 발생한 컨트롤러가 쓰기 작업을 인식하지 못했기 때문에 어떤 경우에도 데이터가 손실되지 않습니다.

사이트 간 연결 실패가 NVRAM 복제를 원격 노드로 차단하는 경우에 더 복잡한 상황이 됩니다. 쓰기가 더 이상 원격 노드에 복제되지 않으므로 컨트롤러에서 심각한 오류가 발생할 경우 데이터가 손실될 수 있습니다. 더 중요한 것은 이러한 상황에서 다른 노드로 페일오버하려고 하면 데이터가 손실된다는 것입니다.

제어 요소는 NVRAM의 동기화 여부입니다. NVRAM이 동기화되면 데이터 손실 위험 없이 노드 간 페일오버를 안전하게 수행할 수 있습니다. MetroCluster 구성에서 NVRAM 및 기본 애그리게이트 플렉스가 동기화되어 있는 경우 데이터 손실의 위험 없이 전환을 진행해도 안전합니다.

ONTAP는 페일오버 또는 스위치오버가 강제 적용되지 않는 한 데이터가 동기화되지 않을 때 페일오버 또는 스위치오버를 허용하지 않습니다. 이러한 방식으로 조건을 강제로 변경하면 데이터가 원래 컨트롤러에 남겨질 수 있으며 데이터 손실이 허용되는 수준임을 알 수 있습니다.

데이터베이스와 기타 애플리케이션은 디스크에 더 큰 내부 데이터 캐시를 유지하기 때문에 페일오버나 스위치오버가 강제 적용되는 경우 손상에 특히 취약합니다. 강제 적용 페일오버 또는 스위치오버가 발생하면 이전에 승인되었던 변경사항이 효과적으로 폐기됩니다. 스토리지 어레이의 콘텐츠가 사실상 이전 시간으로 이동하며, 캐시의 상태는 디스크에 있는 데이터의 상태를 더 이상 반영하지 않습니다.

이러한 상황을 방지하기 위해 ONTAP에서는 NVRAM 장애에 대한 특별 보호를 위해 볼륨을 구성할 수 있습니다. 이 보호 메커니즘이 트리거되면 볼륨이 NVFAIL이라는 상태로 전환됩니다. 이 상태에서는 애플리케이션 충돌을 일으키는 I/O 오류가 발생합니다. 이 충돌로 인해 애플리케이션이 종료되어 오래된 데이터를 사용하지 않습니다. 커밋된 트랜잭션 데이터가 로그에 있어야 하므로 데이터가 손실되지 않아야 합니다. 일반적인 다음 단계는 관리자가 LUN 및 볼륨을 수동으로 다시 온라인 상태로 전환하기 전에 호스트를 완전히 종료하는 것입니다. 이러한 단계에는 일부 작업이 포함될 수 있지만 이 접근 방식은 데이터 무결성을 보장하는 가장 안전한 방법입니다. 모든 데이터에 이 보호가 필요한 것은 아니므로 NVFAIL 동작을 볼륨별로 구성할 수 있습니다.

HA Pair 및 **MetroCluster**

MetroCluster는 2노드 및 HA 쌍의 2가지 구성으로 사용할 수 있습니다. 2노드 구성은 NVRAM에 관한 HA 쌍과 동일하게 동작합니다. 갑작스러운 장애가 발생하는 경우 파트너 노드는 NVRAM 데이터를 재생하여 드라이브의 일관성을 유지하고 확인된 쓰기가 손실되지 않도록 할 수 있습니다.

HA 쌍 구성은 NVRAM을 로컬 파트너 노드에 복제합니다. 컨트롤러 장애가 발생하면 MetroCluster 없이 독립 실행형 HA 쌍을 지원하는 경우처럼, 단순한 컨트롤러 장애가 파트너 노드에서 NVRAM이 재생됩니다. 갑작스러운 전체 사이트 손실이 발생하는 경우 원격 사이트에는 드라이브 일관성을 유지하고 데이터 제공을 시작하는 데 필요한 NVRAM이 있습니다.

MetroCluster의 한 가지 중요한 측면은 원격 노드가 정상적인 운영 조건에서는 파트너 데이터에 액세스할 수 없다는 것입니다. 각 사이트는 기본적으로 반대편의 사이트의 성격을 상정할 수 있는 독립적인 시스템으로서 기능한다. 이 프로세스를 스위치오버라고 하며 계획된 스위치오버를 포함합니다. 사이트 운영이 반대편 사이트로 중단 없이 마이그레이션됩니다. 또한, 재해 복구의 일부로 사이트가 손실되고 수동 또는 자동 전환이 필요한 계획되지 않은 상황이 포함됩니다.

전환 및 스위치백

스위치오버 및 스위치백이란 MetroCluster 구성에서 원격 컨트롤러 간에 볼륨을 전환하는 프로세스를 의미합니다. 이

프로세스는 원격 노드에만 적용됩니다. MetroCluster를 4 볼륨 구성으로 사용할 경우 로컬 노드 페일오버는 앞에서 설명한 테이크오버 및 반환 프로세스가 동일합니다.

계획된 전환 및 스위치백

계획된 스위치오버 또는 스위치백은 노드 간의 테이크오버 또는 반환과 비슷합니다. 이 프로세스에는 여러 단계가 있으며 몇 분이 소요되는 것처럼 보일 수 있지만 실제로 발생하는 것은 스토리지 및 네트워크 리소스의 다중 위상 원활한 전환입니다. 제어 전송이 전체 명령을 실행하는 데 필요한 시간보다 훨씬 빠르게 발생하는 순간입니다.

Takeover/Giveback과 스위치오버/스위치백 간의 주된 차이점은 FC SAN 연결에 영향을 미치는 것입니다. 로컬 테이크오버/반환을 사용하면 호스트에서 로컬 노드에 대한 모든 FC 경로가 손실되고 네이티브 MPIO에 의존하여 사용 가능한 대체 경로로 변경됩니다. 포트는 재배치되지 않습니다. 스위치오버 및 스위치백을 사용하면 컨트롤러의 가상 FC 타겟 포트가 다른 사이트로 전환됩니다. 이러한 애플리케이션은 사실상 SAN에 잠시 존재하지 않게 된 다음 대체 컨트롤러에 다시 나타납니다.

SyncMirror 시간 초과

SyncMirror는 쉘프 장애로부터 보호하는 ONTAP 미러링 기술입니다. 쉘프가 거리를 두고 분리되면 데이터를 원격으로 보호할 수 있습니다.

SyncMirror는 범용 동기식 미러링을 제공하지 않습니다. 결과적으로 가용성이 향상됩니다. 일부 스토리지 시스템은 도미노 모드라고도 하는 일정한 전체 또는 무관 미러링을 사용합니다. 이러한 형태의 미러링은 원격 사이트에 대한 연결이 끊긴 경우 모든 쓰기 작업이 중단되어야 하므로 응용 프로그램에서 제한됩니다. 그렇지 않으면 한 사이트에 쓰기가 존재하지만 다른 사이트에는 쓰기가 존재하지 않습니다. 일반적으로 이러한 환경은 30초 이상 사이트와 사이트 간의 연결이 끊긴 경우 LUN을 오프라인 상태로 전환하도록 구성됩니다.

이 동작은 일부 환경의 하위 집합에 적합합니다. 그러나 대부분의 애플리케이션은 정상적인 작동 조건에서 동기식 복제를 보장하지만 복제를 일시 중지할 수 있는 솔루션이 필요합니다. 사이트 간 연결의 완전한 손실은 주로 재해에 가까운 상황으로 간주됩니다. 일반적으로 이러한 환경은 연결이 복구되거나 데이터 보호를 위해 환경을 종료하기로 결정할 때까지 온라인 상태로 유지되고 데이터를 제공합니다. 순수하게 원격 복제 실패로 인해 애플리케이션을 자동으로 종료해야 하는 요구사항은 특이합니다.

SyncMirror는 시간 초과 방식의 유연성으로 동기식 미러링 요구사항을 지원합니다. 조종기 및/또는 플렉스에 대한 연결이 끊기면 30초 타이머가 카운트 다운을 시작합니다. 카운터가 0에 도달하면 로컬 데이터를 사용하여 쓰기 입출력 처리가 재개됩니다. 데이터의 원격 복제본을 사용할 수 있지만 연결이 복원될 때까지 시간이 지나면 동결됩니다. 재동기화는 애그리게이트 레벨 스냅샷을 활용하여 가능한 한 빨리 시스템을 동기식 모드로 되돌립니다.

특히 대부분의 경우 이러한 종류의 범용 전체 또는 무관 도미노 모드 복제는 애플리케이션 계층에서 더 잘 구현됩니다. 예를 들어 Oracle DataGuard에는 모든 상황에서 장기 인스턴스 복제를 보장하는 최대 보호 모드가 포함되어 있습니다. 구성 가능한 시간 제한을 초과하는 기간 동안 복제 링크가 실패하면 데이터베이스가 종료됩니다.

패브릭 연결 **MetroCluster**를 통한 자동 무인 전환

자동 무인 전환(AUSO)은 크로스 사이트 HA의 형태를 제공하는 패브릭 연결 MetroCluster 기능입니다. 앞서 설명했듯이, MetroCluster는 각 사이트의 단일 컨트롤러 또는 각 사이트의 HA 쌍 두 가지로 사용할 수 있습니다. HA 옵션의 주요 이점은 계획되었거나 계획되지 않은 컨트롤러 종료를 통해 모든 I/O를 로컬에 둘 수 있다는 것입니다. 단일 노드 옵션의 이점은 비용, 복잡성 및 인프라의 감소입니다.

AUSO의 주요 가치는 Fabric Attached MetroCluster 시스템의 HA 기능을 개선하는 것입니다. 각 사이트가 반대 사이트의 상태를 모니터링하며, 데이터를 제공할 노드가 남아 있지 않으면 AUSO로 인해 빠른 전환이 발생합니다. 이 접근 방식은 가용성 측면에서 구성이 HA 쌍에 더 가깝게 배치되기 때문에 사이트당 단일 노드만을 사용하는 MetroCluster 구성에서 특히 유용합니다.

AUSO는 HA 쌍 수준에서 포괄적인 모니터링을 제공할 수 없습니다. HA 쌍은 노드 간 직접 통신을 위한 이중화 물리적 케이블 2개가 포함되어 있기 때문에 매우 높은 가용성을 제공할 수 있습니다. 또한 HA 쌍의 두 노드는 이중 루프의 동일한 디스크 세트에 액세스할 수 있어, 한 노드에서 다른 노드의 상태를 모니터링할 수 있는 또 다른 경로를 제공합니다.

MetroCluster 클러스터는 사이트 간 네트워크 연결을 통해 노드 간 통신과 디스크 액세스가 모두 필요한 사이트 전체에 존재합니다. 클러스터의 나머지 하트비트를 모니터링하는 기능은 제한되어 있습니다. AUSO는 네트워크 문제로 인해 다른 사이트가 사용할 수 없는 상황이 아니라 실제로 다운된 상황을 구분해야 합니다.

그 결과, HA 쌍의 컨트롤러에서 시스템 패닉 같은 특정 이유로 컨트롤러 장애를 감지하면 테이크오버를 프롬프트 상태가 될 수 있습니다. 또한 하트비트 손실이라고도 하는 연결이 완전히 끊긴 경우 Takeover를 프롬프트할 수도 있습니다.

MetroCluster 시스템은 원래 사이트에서 특정 장애가 감지되는 경우에만 자동 전환을 안전하게 수행할 수 있습니다. 또한 스토리지 시스템의 소유권을 가져오는 컨트롤러는 디스크 및 NVRAM 데이터의 동기화를 보장할 수 있어야 합니다. 컨트롤러는 여전히 작동 가능한 소스 사이트와의 접촉이 끊겼다는 이유로 스위치오버의 안전을 보장할 수 없습니다. 스위치오버 자동화를 위한 추가 옵션은 다음 섹션에서 MetroCluster Tiebreaker(MCTB) 솔루션에 관한 정보를 참조하십시오.

패브릭 연결 **MetroCluster**가 포함된 **MetroCluster Tiebreaker**

를 클릭합니다 "NetApp MetroCluster Tiebreaker의 약어입니다" 소프트웨어를 세 번째 사이트에서 실행하여 MetroCluster 환경의 상태를 모니터링하고, 알림을 보내고, 재해 상황에서 선택적으로 스위치오버를 수행할 수 있습니다. 타이브레이커에 대한 자세한 설명은 에서 확인할 수 있습니다 "NetApp Support 사이트"하지만 MetroCluster Tiebreaker의 주요 목적은 사이트 손실을 감지하는 것입니다. 또한 사이트 손실과 연결 손실 간에 구분해야 합니다. 예를 들어, Tiebreaker가 운영 사이트에 연결할 수 없기 때문에 전환이 발생하지 않아야 합니다. 따라서 Tiebreaker는 원격 사이트의 운영 사이트 접속 기능을 모니터링합니다.

AUSO를 통한 자동 절체는 MCTB와도 호환됩니다. AUSO는 특정 장애 이벤트를 감지한 다음 NVRAM 및 SyncMirror 플렉스가 동기화되어 있는 경우에만 스위치오버를 호출하도록 설계되었기 때문에 매우 빠르게 대응합니다.

반대로 타이브레이커는 원격으로 위치하므로 타이머가 경과할 때까지 기다린 후 사이트를 비활성화해야 합니다. Tiebreaker는 결국 AUSO에 포함된 일종의 컨트롤러 장애를 감지하지만, 일반적으로 AUSO는 이미 전환을 시작하고 Tiebreaker가 작동하기 전에 전환을 완료했을 수 있습니다. Tiebreaker에서 생성된 두 번째 switchover 명령은 거부됩니다.

- 주의: * MCTB 소프트웨어는 전환을 강제 적용할 때 NVRAM이 동기화되었는지 또는 플렉스가 동기화되었는지 확인하지 않습니다. 자동 전환이 구성된 경우 유지 관리 활동 중에 NVRAM 또는 SyncMirror 플렉스의 동기화가 손실되는 것을 방지해야 합니다.

또한 MCTB는 지속적인 재해를 처리하지 못해 다음과 같은 일련의 이벤트가 발생할 수 있습니다.

1. 사이트 간 연결이 30초 이상 중단됩니다.
2. SyncMirror 복제 시간이 초과되고 운영 사이트에서 작업이 계속되어 원격 복제본이 오래된 상태로 남습니다.
3. 기본 사이트가 손실되어 기본 사이트에 복제되지 않은 변경 내용이 있습니다. 이렇게 하면 다음과 같은 여러 가지 이유로 전환이 바람직하지 않을 수 있습니다.
    - 기본 사이트에 중요 데이터가 있을 수 있으며 이 경우 해당 데이터를 복구할 수 있습니다. 애플리케이션의 지속적인 운영을 허용한 전환은 중요 데이터를 효과적으로 폐기합니다.
    - 사이트 손실 시 기본 사이트의 스토리지 리소스를 사용 중이던 정상적인 사이트의 애플리케이션이 데이터를 캐싱했을 수 있습니다. 스위치오버로 인해 캐시와 일치하지 않는 오래된 데이터가 생성됩니다.
    - 사이트 손실 시 기본 사이트의 스토리지 리소스를 사용하고 있었던 정상적인 사이트의 운영 체제에서는 데이터가 캐시되었을 수 있습니다. 스위치오버로 인해 캐시와 일치하지 않는 오래된 데이터가 생성됩니다. 가장

안전한 옵션은 사이트 장애가 감지되면 알림을 보내도록 Tiebreaker를 구성한 다음 사람이 전환을 강제 적용할 것인지 여부를 결정하도록 하는 것입니다. 캐시된 데이터를 지우려면 먼저 응용 프로그램 및/또는 운영 체제를 종료해야 할 수 있습니다. 또한 NVFAIL 설정을 사용하여 보호 기능을 추가하고 장애 조치 프로세스를 간소화할 수 있습니다.

MetroCluster IP를 사용하는 **ONTAP** 중재자

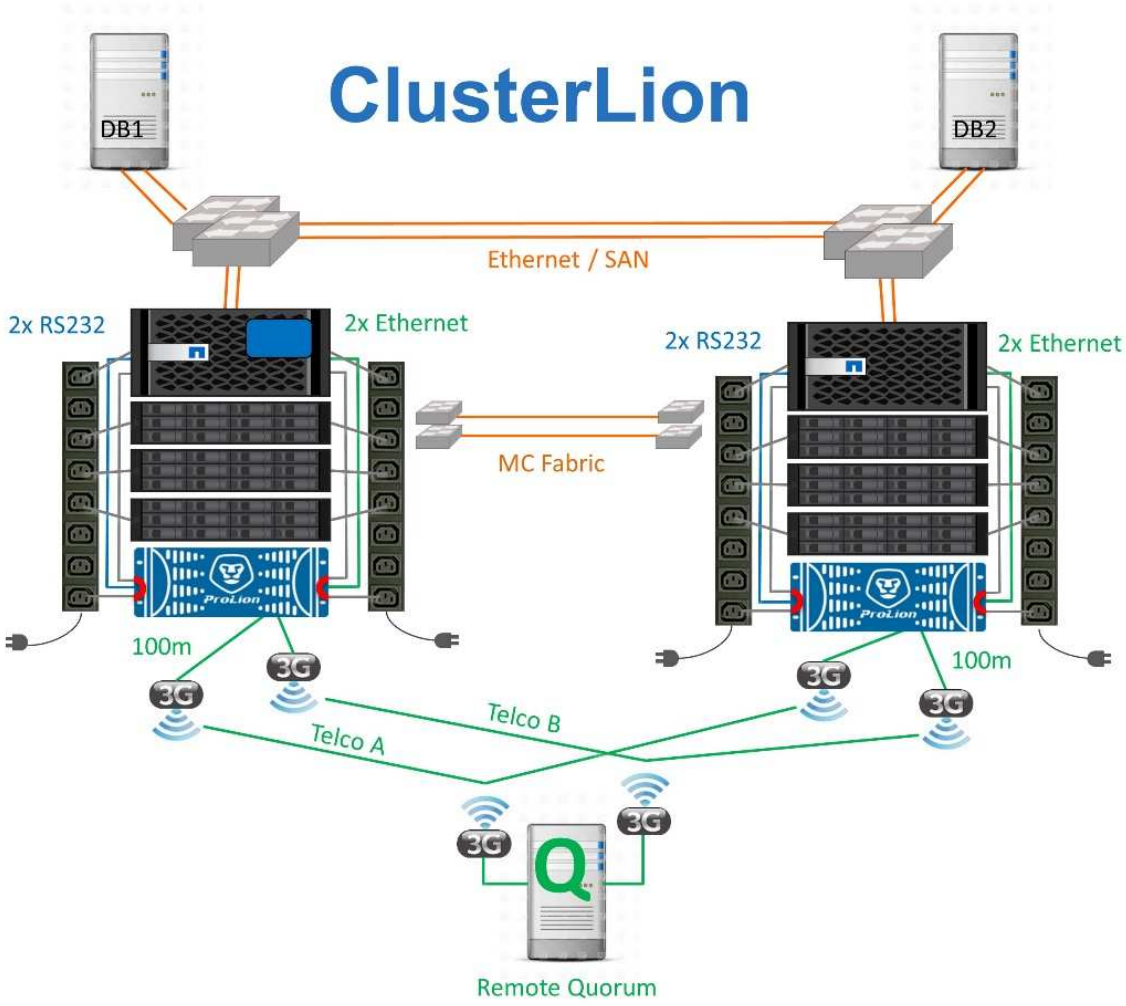
ONTAP mediator는 MetroCluster IP 및 기타 특정 ONTAP 솔루션과 함께 사용됩니다. 위에서 설명한 MetroCluster Tiebreaker 소프트웨어와 마찬가지로 기존 Tiebreaker 서비스 역할을 하지만 자동 자동 전환을 수행하는 중요한 기능도 포함되어 있습니다.

패브릭이 연결된 MetroCluster는 반대쪽 사이트의 스토리지 장치에 직접 액세스할 수 있습니다. 이를 통해 한 MetroCluster 컨트롤러가 드라이브에서 하트비트 데이터를 읽어 다른 컨트롤러의 상태를 모니터링할 수 있습니다. 이를 통해 한 컨트롤러가 다른 컨트롤러의 장애를 인식하고 전환을 수행할 수 있습니다.

반면, MetroCluster IP 아키텍처는 컨트롤러-컨트롤러 연결을 통해서만 모든 I/O를 라우팅하며, 원격 사이트의 스토리지 장치에 직접 액세스할 수 없습니다. 이로 인해 컨트롤러가 장애를 감지하고 스위치오버를 수행할 수 없게 됩니다. 따라서 사이트 손실을 감지하고 자동으로 전환을 수행하기 위한 Tiebreaker 장치로 ONTAP 중재자가 필요합니다.

ClusterLion이 포함된 가상 **3**번째 사이트

ClusterLion은 가상 3차 사이트로 작동하는 고급 MetroCluster 모니터링 어플라이언스입니다. 이 접근 방식을 통해 MetroCluster는 완전 자동화된 스위치오버 기능을 통해 2개 사이트 구성으로 안전하게 구축할 수 있습니다. 또한 ClusterLion은 추가 네트워크 수준 모니터를 수행하고 전환 후 작업을 실행할 수 있습니다. ProLion에서 전체 문서를 다운로드할 수 있습니다.



- ClusterLion 어플라이언스는 이더넷 및 직렬 케이블을 직접 연결하여 컨트롤러의 상태를 모니터링합니다.
- 이 두 장비는 이중화 3G 무선 연결을 통해 서로 연결됩니다.
- ONTAP 컨트롤러의 전원은 내부 릴레이를 통해 배선됩니다. 사이트 장애가 발생할 경우 내부 UPS 시스템이 포함된 ClusterLion은 전환을 호출하기 전에 전원 연결을 끊습니다. 이 과정을 통해 브레인 분할 상태가 발생하지 않도록 합니다.
- ClusterLion은 30초 SyncMirror 타임아웃 내에 전환을 수행하거나 전혀 전환하지 않습니다.
- NVRAM 및 SyncMirror 플렉스의 상태가 동기화되어 있지 않으면 ClusterLion은 전환을 수행하지 않습니다.
- ClusterLion은 MetroCluster가 완전히 동기화된 경우에만 전환을 수행하기 때문에 NVFAIL은 필요하지 않습니다. 이렇게 구성하면 확장된 Oracle RAC와 같은 사이트 확장 환경이 계획되지 않은 전환 중에도 온라인 상태를 유지할 수 있습니다.
- 여기에는 패브릭 연결 MetroCluster 및 MetroCluster IP가 모두 포함됩니다

Oracle 데이터베이스 및 SyncMirror

MetroCluster 시스템을 통한 Oracle 데이터 보호의 기반은 최대 성능 스케일아웃 동기식 미러링 기술인 SyncMirror입니다.

SyncMirror를 사용한 데이터 보호

가장 간단한 수준인 동기식 복제는 미러링된 스토리지의 양쪽에서 변경 사항이 확인되기 전에 수행되어야 함을 의미합니다. 예를 들어, 데이터베이스에서 로그를 작성하거나 VMware 게스트에 패치를 적용하는 경우 쓰기가 손실되지 않아야 합니다. 프로토콜 레벨에서 스토리지 시스템은 두 사이트의 비휘발성 미디어에 커밋될 때까지 쓰기를 인증해서는 안 됩니다. 그래야만 데이터 손실의 위험 없이 진행하는 것이 안전합니다.

동기식 복제 솔루션을 설계하고 관리하는 첫 번째 단계는 동기식 복제 기술을 사용하는 것입니다. 가장 중요한 고려 사항은 계획된 고장 시나리오와 예상치 못한 다양한 장애 시나리오 중에 발생할 수 있는 상황을 이해하는 것입니다. 모든 동기식 복제 솔루션이 동일한 기능을 제공하는 것은 아닙니다. 데이터 손실이 0인 복구 지점 목표(RPO)를 제공하는 솔루션이 필요한 경우 모든 장애 시나리오를 고려해야 합니다. 특히 사이트 간 연결 손실로 인해 복제가 불가능할 때 예상되는 결과는 무엇입니까?

SyncMirror 데이터 가용성

MetroCluster 복제는 NetApp SyncMirror 기술을 기반으로 하며 동기식 모드로 효율적으로 전환하거나 아웃하도록 설계되었습니다. 이 기능은 동기식 복제를 필요로 하지만 데이터 서비스를 위해 고가용성이 필요한 고객의 요구사항을 충족합니다. 예를 들어 원격 사이트에 대한 연결이 끊어진 경우 일반적으로 스토리지 시스템이 복제되지 않은 상태로 계속 작동하도록 하는 것이 좋습니다.

대부분의 동기식 복제 솔루션은 동기식 모드에서만 작동할 수 있습니다. 이러한 유형의 모든 또는 무관 복제를 도미노 모드라고도 합니다. 이러한 스토리지 시스템은 데이터의 로컬 및 원격 복제본이 동기화되지 않도록 하는 대신 데이터 제공을 중지합니다. 복제가 강제로 중단되면 재동기화는 시간이 매우 오래 걸리고 미러링이 다시 설정되는 동안 고객이 완전한 데이터 손실에 노출되도록 할 수 있습니다.

원격 사이트에 연결할 수 없는 경우 SyncMirror가 동기식 모드를 원활하게 전환할 수 있을 뿐만 아니라 연결이 복원되면 RPO=0 상태로 빠르게 다시 동기화할 수 있습니다. 또한 재동기화 중에 원격 사이트의 오래된 데이터 복제본을 사용 가능한 상태로 유지할 수 있으므로 데이터의 로컬 및 원격 복제본이 항상 존재합니다.

도미노 모드가 필요한 경우 NetApp은 SnapMirror Synchronous(SM-S)를 제공합니다. Oracle DataGuard 또는 호스트측 디스크 미러링의 시간 초과와 같은 애플리케이션 레벨 옵션도 있습니다. 자세한 정보와 옵션은 NetApp 또는 파트너 계정 팀에 문의하십시오.

MetroCluster을 사용한 Oracle 데이터베이스 페일오버

```
 Metrocluster is an ONTAP feature that can protect your Oracle databases
with RPO=0 synchronous mirroring across sites, and it scales up to support
hundreds of databases on a single MetroCluster system. It's also simple to
use. The use of MetroCluster does not necessarily add to or change any
best practices for operating a enterprise applications and databases.
```

일반적인 모범 사례가 여전히 적용되므로 필요한 경우 `RPO = 0` 데이터 보호만 있으면 `MetroCluster`를 통해 이 요구사항을 충족할 수 있습니다. 하지만 대부분의 고객은 `RPO=0` 데이터 보호를 위해뿐만 아니라 재해 시나리오 중에 `RTO`를 개선하고 사이트 유지 관리 작업의 일부로 투명한 페일오버를 제공하기 위해 `MetroCluster`을 사용합니다.

사전 구성된 OS로 페일오버

SyncMirror은 재해 복구 사이트에서 데이터의 동기식 복사본을 제공하지만, 데이터를 사용하려면 운영 체제와 관련 애플리케이션이 필요합니다. 기본 자동화를 통해 전체 환경의 장애 조치 시간을 크게 개선할 수 있습니다. Oracle RAC,

VCS(Veritas Cluster Server) 또는 VMware HA 같은 Clusterware 제품은 사이트 전체에 클러스터를 생성하는 데 자주 사용되며, 대부분의 경우 간단한 스크립트로 페일오버 프로세스를 구동할 수 있습니다.

운영 노드가 손실되면 대체 사이트에서 애플리케이션을 온라인으로 전환하도록 클러스터웨어(또는 스크립트)가 구성됩니다. 한 가지 옵션은 애플리케이션을 구성하는 NFS 또는 SAN 리소스에 대해 사전 구성된 대기 서버를 생성하는 것입니다. 운영 사이트에 장애가 발생하면 클러스터웨어 또는 스크립트된 대체 시스템이 다음과 유사한 일련의 작업을 수행합니다.

1. MetroCluster 강제 전환
2. FC LUN 검색 수행(SAN만 해당)
3. 파일 시스템을 마운트하는 중입니다
4. 응용 프로그램을 시작하는 중입니다

이 방법의 주요 요구 사항은 원격 사이트에서 실행 중인 OS입니다. 애플리케이션 바이너리로 사전 구성되어야 합니다. 즉, 패치와 같은 작업은 운영 및 대기 사이트에서 수행되어야 합니다. 또는 재해가 선언된 경우 애플리케이션 바이너리를 원격 사이트로 미러링하고 마운트할 수 있습니다.

실제 활성화 절차는 간단합니다. LUN 검색과 같은 명령은 FC 포트당 몇 개의 명령만 사용하면 됩니다. 파일 시스템 마운팅은 에 불과합니다 `mount` CLI에서 단일 명령으로 명령 및 데이터베이스와 ASM을 모두 시작하고 중지할 수 있습니다. 볼륨 및 파일 시스템이 전환 전 재해 복구 사이트에서 사용되지 않는 경우에는 설정할 필요가 없습니다 `dr-force- nvfail` On 볼륨.

가상화된 **OS**로 페일오버

데이터베이스 환경의 페일오버는 운영 체제 자체를 포함하도록 확장할 수 있습니다. 이론적으로 이 페일오버는 부팅 LUN에서 수행할 수 있지만 대부분의 경우 가상화된 OS에서 수행됩니다. 절차는 다음 단계와 유사합니다.

1. MetroCluster 강제 전환
2. 데이터베이스 서버 가상 머신을 호스팅하는 데이터 저장소를 마운트합니다
3. 가상 머신 시작
4. 데이터베이스를 수동으로 시작하거나 가상 시스템이 데이터베이스를 자동으로 시작하도록 구성합니다

예를 들어, ESX 클러스터가 사이트에 걸쳐 있을 수 있습니다. 재해 발생 시 전환 후 재해 복구 사이트에서 가상 시스템을 온라인으로 전환할 수 있습니다. 재해 발생 시 가상 데이터베이스 서버를 호스팅하는 데이터 저장소를 사용하지 않는 한 설정할 필요가 없습니다 `dr-force- nvfail` 연결된 볼륨에서.

모든 주요 **MetroCluster** 파이버 채널 인프라 환경을 지원합니다

NVFAIL은 데이터베이스를 통해 데이터 무결성 보호를 극대화하도록 설계된 ONTAP의 일반적인 데이터 무결성 기능입니다.

이 섹션에서는 MetroCluster 관련 주제를 다루는 기본 ONTAP NVFAIL에 대한 설명을 확장합니다.

MetroCluster를 사용할 경우, 쓰기가 하나 이상의 다른 컨트롤러의 로컬 NVRAM 및 NVRAM에 로그인되기 전까지는 승인되지 않습니다. 이렇게 하면 하드웨어 장애나 정전이 발생해도 전송 중인 I/O가 손실되지 않습니다 로컬 NVRAM에 장애가 발생하거나 다른 노드에 대한 연결이 실패하면 데이터가 더 이상 미러링되지 않습니다.

로컬 NVRAM에 오류가 보고되면 노드가 종료됩니다. 이 종료를 통해 HA Pair를 사용할 경우 파트너 컨트롤러로

페일오버됩니다. MetroCluster를 사용할 경우 선택한 전체 구성에 따라 동작이 달라지지만 원격 메모로 자동 페일오버될 수 있습니다. 오류가 발생한 컨트롤러가 쓰기 작업을 인식하지 못했기 때문에 어떤 경우에도 데이터가 손실되지 않습니다.

사이트 간 연결 실패가 NVRAM 복제를 원격 노드로 차단하는 경우에 더 복잡한 상황이 됩니다. 쓰기가 더 이상 원격 노드에 복제되지 않으므로 컨트롤러에서 심각한 오류가 발생할 경우 데이터가 손실될 수 있습니다. 더 중요한 것은 이러한 상황에서 다른 노드로 페일오버하려고 하면 데이터가 손실된다는 것입니다.

제어 요소는 NVRAM의 동기화 여부입니다. NVRAM이 동기화되면 데이터 손실 위험 없이 노드 간 페일오버를 안전하게 수행할 수 있습니다. MetroCluster 구성에서 NVRAM 및 기본 애그리게이트 플렉스가 동기화되어 있는 경우 데이터 손실 위험 없이 전환을 진행해도 안전합니다.

ONTAP는 페일오버 또는 스위치오버가 강제 적용되지 않는 한 데이터가 동기화되지 않을 때 페일오버 또는 스위치오버를 허용하지 않습니다. 이러한 방식으로 조건을 강제로 변경하면 데이터가 원래 컨트롤러에 남겨질 수 있으며 데이터 손실이 허용되는 수준임을 알 수 있습니다.

데이터베이스는 디스크에 더 큰 내부 데이터 캐시를 유지하기 때문에 페일오버나 스위치오버가 강제 적용되는 경우 손상에 특히 취약합니다. 강제 적용 페일오버 또는 스위치오버가 발생하면 이전에 승인되었던 변경사항이 효과적으로 폐기됩니다. 스토리지 어레이의 콘텐츠가 사실상 이전 시간으로 이동하며, 데이터베이스 캐시의 상태는 디스크에 있는 데이터의 상태를 더 이상 반영하지 않습니다.

이 상황에서 애플리케이션을 보호하기 위해 ONTAP에서는 NVRAM 장애에 대비하여 특별한 보호를 제공하도록 볼륨을 구성할 수 있습니다. 이 보호 메커니즘이 트리거되면 볼륨이 NVFAIL이라는 상태로 전환됩니다. 이 상태에서는 애플리케이션 종료가 I/O 오류가 발생하여 오래된 데이터를 사용하지 않습니다. 확인된 쓰기가 스토리지 시스템에 계속 존재하고 데이터베이스의 경우 커밋된 트랜잭션 데이터가 로그에 있어야 하므로 데이터가 손실되지 않아야 합니다.

일반적인 다음 단계는 관리자가 LUN 및 볼륨을 수동으로 다시 온라인 상태로 전환하기 전에 호스트를 완전히 종료하는 것입니다. 이러한 단계에는 일부 작업이 포함될 수 있지만 이 접근 방식은 데이터 무결성을 보장하는 가장 안전한 방법입니다. 모든 데이터에 이 보호가 필요한 것은 아니므로 NVFAIL 동작을 볼륨별로 구성할 수 있습니다.

수동으로 **NVFAIL**을 강제 적용합니다

사이트 전체에 분산된 애플리케이션 클러스터(VMware, Oracle RAC 등 포함)를 사용하여 강제 전환할 수 있는 가장 안전한 옵션은 을 지정하는 것입니다 `-force-nvfail-all` 명령줄에 입력합니다. 이 옵션은 캐시된 모든 데이터를 플러시하기 위한 긴급 조치로 사용할 수 있습니다. 호스트에서 원래 재해 복구 사이트에 있는 스토리지 리소스를 사용하는 경우 입출력 오류 또는 오래된 파일 핸들이 발생합니다 (`ESTALE`) 오류. Oracle 데이터베이스가 충돌하고 파일 시스템이 완전히 오프라인 상태가 되거나 읽기 전용 모드로 전환됩니다.

전환이 완료된 후 은 을(를) 수행합니다 `in-nvfailed-state` 플래그를 지워야 하며 LUN을 온라인 상태로 설정해야 합니다. 이 작업이 완료되면 데이터베이스를 다시 시작할 수 있습니다. 이러한 작업을 자동화하여 RTO를 줄일 수 있습니다.

dr-force-nvfail입니다

일반적인 안전 조치로 을 설정합니다 `dr-force-nvfail` 정상 작업 중에 원격 사이트에서 액세스할 수 있는 모든 볼륨에 플래그를 표시하므로, 페일오버 전에 사용된 활동입니다. 이 설정의 결과로 선택한 원격 볼륨이 들어가면 사용할 수 없게 됩니다 `in-nvfailed-state` 스위치오버 중에 전환이 완료된 후 은 을(를) 수행합니다 `in-nvfailed-state` 플래그를 지워야 하며 LUN을 온라인 상태로 설정해야 합니다. 이러한 작업이 완료되면 응용 프로그램을 다시 시작할 수 있습니다. 이러한 작업을 자동화하여 RTO를 줄일 수 있습니다.

결과는 을 사용하는 것과 같습니다 `-force-nvfail-all` 수동 전환 플래그 그러나 영향을 받는 볼륨의 수는 오래된 캐시가 있는 애플리케이션이나 운영 체제에서 보호되어야 하는 볼륨으로만 제한될 수 있습니다.

을 사용하지 않는 환경에는 두 가지 중요한 요구사항이 있습니다 `dr-force-nvfail` 애플리케이션 볼륨에서:

- 강제 적용 스위치오버는 1차 사이트 손실 후 30초 이내여야 합니다.
- 유지보수 작업 중 또는 SyncMirror 플렉스 또는 NVRAM 복제가 동기화되지 않는 기타 조건에서는 전환이 발생하지 않아야 합니다. 사이트 장애 발생 후 30초 이내에 전환을 수행하도록 구성된 Tiebreaker 소프트웨어를 사용하여 첫 번째 요구사항을 충족할 수 있습니다. 이 요구사항이 사이트 장애 감지 후 30초 이내에 전환을 수행해야 함을 의미하는 것은 아닙니다. 즉, 사이트가 작동 가능으로 확인된 후 30초가 경과하면 강제로 전환을 수행하는 것이 더 이상 안전하지 않습니다.

MetroCluster 구성이 동기화되지 않은 것으로 알려진 경우 모든 자동 전환 기능을 비활성화하여 두 번째 요구 사항을 부분적으로 충족할 수 있습니다. 더 좋은 옵션은 NVRAM 복제 및 SyncMirror Plex의 상태를 모니터링할 수 있는 Tiebreaker 솔루션을 구축하는 것입니다. 클러스터가 완전히 동기화되지 않은 경우 Tiebreaker가 전환을 트리거해서는 안 됩니다.

NetApp MCTB 소프트웨어는 동기화 상태를 모니터링할 수 없으므로 어떤 이유로든 MetroCluster가 동기화되지 않은 경우 이 기능을 비활성화해야 합니다. ClusterLion에는 NVRAM 모니터링 및 플렉스 모니터링 기능이 포함되어 있으며, MetroCluster 시스템이 완전히 동기화되는 것으로 확인되지 않는 한 전환을 트리거하지 않도록 구성할 수 있습니다.

MetroCluster에 있는 Oracle 단일 인스턴스

앞서 설명한 것처럼 MetroCluster 시스템이 있다고 해서 데이터베이스 운영에 대한 모범 사례가 반드시 추가되지 않거나 변경되는 것은 아닙니다. 고객 MetroCluster 시스템에서 현재 실행 중인 데이터베이스의 대부분은 단일 인스턴스이며 Oracle on ONTAP 설명서의 권장 사항을 따릅니다.

사전 구성된 OS로 페일오버

SyncMirror은 재해 복구 사이트에서 데이터의 동기식 복사본을 제공하지만, 데이터를 사용하려면 운영 체제와 관련 애플리케이션이 필요합니다. 기본 자동화를 통해 전체 환경의 장애 조치 시간을 크게 개선할 수 있습니다. VCS(Veritas Cluster Server)와 같은 클러스터웨어 제품은 사이트 전체에 클러스터를 생성하는 데 자주 사용되며, 대부분의 경우 간단한 스크립트로 페일오버 프로세스를 구동할 수 있습니다.

운영 노드가 손실되면 대체 사이트에서 데이터베이스를 온라인으로 전환하도록 클러스터웨어(또는 스크립트)가 구성됩니다. 한 가지 옵션은 데이터베이스를 구성하는 NFS 또는 SAN 리소스에 대해 사전 구성된 대기 서버를 생성하는 것입니다. 운영 사이트에 장애가 발생하면 클러스터웨어 또는 스크립트된 대체 시스템이 다음과 유사한 일련의 작업을 수행합니다.

1. MetroCluster 강제 전환
2. FC LUN 검색 수행(SAN만 해당)
3. 파일 시스템 마운트 및/또는 ASM 디스크 그룹 마운트
4. 데이터베이스를 시작하는 중입니다

이 방법의 주요 요구 사항은 원격 사이트에서 실행 중인 OS입니다. Oracle 바이너리로 사전 구성되어야 합니다. 즉, Oracle 패치 적용과 같은 작업이 운영 및 대기 사이트에서 수행되어야 합니다. 또는 재해가 선언된 경우 Oracle 바이너리를 원격 사이트로 미러링하고 마운트할 수 있습니다.

실제 활성화 절차는 간단합니다. LUN 검색과 같은 명령은 FC 포트당 몇 개의 명령만 사용하면 됩니다. 파일 시스템 마운팅은 에 불과합니다 `mount` CLI에서 단일 명령으로 명령 및 데이터베이스와 ASM을 모두 시작하고 중지할 수 있습니다. 볼륨 및 파일 시스템이 전환 전 재해 복구 사이트에서 사용되지 않는 경우에는 설정할 필요가 없습니다 `dr-`

`force- nvfail` On 볼륨.

가상화된 OS로 페일오버

데이터베이스 환경의 페일오버는 운영 체제 자체를 포함하도록 확장할 수 있습니다. 이론적으로 이 페일오버는 부팅 LUN에서 수행할 수 있지만 대부분의 경우 가상화된 OS에서 수행됩니다. 절차는 다음 단계와 유사합니다.

1. MetroCluster 강제 전환
2. 데이터베이스 서버 가상 머신을 호스팅하는 데이터 저장소를 마운트합니다
3. 가상 머신 시작
4. 데이터베이스를 수동으로 시작하거나 데이터베이스를 자동으로 시작하도록 가상 시스템을 구성하면 ESX 클러스터가 사이트에 걸쳐 있을 수 있습니다. 재해 발생 시 전환 후 재해 복구 사이트에서 가상 시스템을 온라인으로 전환할 수 있습니다. 재해 발생 시 가상 데이터베이스 서버를 호스팅하는 데이터 저장소를 사용하지 않는 한 설정할 필요가 없습니다 `dr-force- nvfail` 연결된 볼륨에서.

MetroCluster에서 Oracle RAC 확장

많은 고객이 사이트 간에 Oracle RAC 클러스터를 확장하여 완벽한 Active-Active 구성을 실현함으로써 RTO를 최적화합니다. Oracle RAC의 쿼럼 관리를 포함해야 하기 때문에 전체 설계가 더 복잡해집니다. 또한, 두 사이트에서 데이터에 액세스할 수 있으므로 강제 전환으로 인해 최신 데이터 복사본이 사용될 수 있습니다.

두 사이트 모두에 데이터 복사본이 있지만 현재 애그리게이트를 소유하고 있는 컨트롤러만 데이터를 제공할 수 있습니다. 따라서 확장된 RAC 클러스터의 경우 원격 노드가 사이트 간 연결에서 I/O를 수행해야 합니다. 결과적으로 I/O 지연 시간이 추가되지만 이 지연 시간은 일반적으로 문제가 되지 않습니다. RAC 상호 연결 네트워크도 사이트 간에 확장해야 하므로 지연 시간이 짧은 고속 네트워크가 필요합니다. 추가된 지연 시간으로 인해 문제가 발생할 경우 클러스터를 액티브-패시브 방식으로 작동할 수 있습니다. 그런 다음 I/O 집약적인 작업을 애그리게이트가 속한 컨트롤러에 로컬인 RAC 노드로 보내야 함. 그런 다음 원격 노드가 가벼운 I/O 작업을 수행하거나 온전한 대기 서버로만 사용됩니다.

액티브-액티브 확장 RAC가 필요한 경우 MetroCluster 대신 ASM 미러링을 고려해야 합니다. ASM 미러링을 사용하면 데이터의 특정 복제본을 선호할 수 있습니다. 따라서 모든 읽기가 로컬에서 실행되는 확장 RAC 클러스터를 구축할 수 있습니다. 읽기 I/O가 사이트를 통과하지 않으므로 지연 시간이 가장 짧습니다. 모든 쓰기 작업은 사이트 간 연결을 전송해야 하지만 동기식 미러링 솔루션에서 이러한 트래픽은 피할 수 없습니다.

> 가상화된 부팅 디스크를 비롯한 부팅 LUN을 Oracle RAC와 함께 사용하는 경우, 이 명령을 사용합니다 `misscount` 매개 변수를 변경해야 할 수 있습니다. RAC 시간 초과 매개변수에 대한 자세한 내용은 을 참조하십시오 "ONTAP 지원 Oracle RAC".

2개 사이트 구성

2개 사이트의 확장 RAC 구성은 운영 중단 없이 많은 재해 시나리오에서도 가동 중단 없이 지속되는 액티브-액티브 데이터베이스 서비스를 제공할 수 있습니다.

RAC 보팅 파일

MetroCluster에서 확장 RAC를 구축할 때 가장 먼저 고려해야 할 사항은 쿼럼 관리입니다. Oracle RAC에는 디스크 하트비트와 네트워크 하트비트를 관리하는 두 가지 메커니즘이 있습니다. 디스크 하트비트는 보팅 파일을 사용하여

스토리지 액세스를 모니터링합니다. 단일 사이트 RAC 구성의 경우 기본 스토리지 시스템이 HA 기능을 제공하는 한 단일 보팅 리소스로 충분합니다.

이전 버전의 Oracle에서는 보팅 파일이 물리적 스토리지 장치에 배치되었지만 현재 버전의 Oracle에서는 보팅 파일이 ASM 디스크 그룹에 저장됩니다.

> Oracle RAC는 NFS에서 지원됩니다. 그리드 설치 프로세스 중에 그리드 파일에 사용되는 NFS 위치를 ASM 디스크 그룹으로 제공하기 위한 일련의 ASM 프로세스가 생성됩니다. 이 프로세스는 최종 사용자에게 거의 투명하며 설치가 완료된 후 지속적인 ASM 관리가 필요하지 않습니다.

2개 사이트 구성의 첫 번째 요구 사항은 무중단 재해 복구 프로세스를 보장하는 방식으로 각 사이트에서 투표 파일의 절반 이상을 항상 액세스할 수 있도록 하는 것입니다. 이 작업은 투표 파일이 ASM 디스크 그룹에 저장되기 전에는 간단했지만, 오늘날 관리자는 ASM 중복의 기본 원칙을 이해해야 합니다.

ASM 디스크 그룹에는 이중화를 위한 세 가지 옵션이 있습니다 `external`, `normal`, 및 `high`. 즉, 미러링되지 않은, 미러링된, 3웨이 미러링이 있습니다. 라는 새로운 옵션입니다 `Flex` 사용 가능하지만 거의 사용되지 않습니다. 이중화 수준 및 중복 장치의 배치가 장애 시나리오에서 수행되는 작업을 제어합니다. 예를 들면 다음과 같습니다.

- 에 투표 파일 배치 `diskgroup` 와 함께 `external` 이중화 리소스는 사이트 간 연결이 끊긴 경우 하나의 사이트를 제거할 수 있도록 보장합니다.
- 에 투표 파일 배치 `diskgroup` 와 함께 `normal` 사이트당 하나의 ASM 디스크만 있는 중복으로 인해 두 사이트 간 연결이 끊어지면 두 사이트 모두에서 노드 제거가 보장됩니다.
- 에 투표 파일 배치 `diskgroup` 와 함께 `high` 한 사이트에 두 개의 디스크가 있고 다른 사이트에 한 개의 디스크가 있는 중복성을 통해 두 사이트가 모두 작동 중이고 상호 연결할 수 있는 경우 활성-활성 작업이 가능합니다. 그러나 단일 디스크 사이트가 네트워크에서 분리되어 있으면 해당 사이트가 제거됩니다.

RAC 네트워크 하트비트

Oracle RAC 네트워크 하트비트는 클러스터 상호 연결에서 노드 가용성을 모니터링합니다. 클러스터에 남아 있으려면 노드가 다른 노드의 절반 이상에 연결할 수 있어야 합니다. 2개 사이트 아키텍처에서는 이 요구 사항으로 인해 RAC 노드 수를 다음과 같이 선택할 수 있습니다.

- 사이트당 동일한 수의 노드를 배치하면 네트워크 연결이 끊어질 경우 한 사이트에서 제거됩니다.
- 한 사이트에 N 노드를 배치하고 반대쪽 사이트에 N+1 노드를 배치하면 사이트 간 연결이 끊어지면 사이트가 네트워크 쿼럼에 더 많은 노드를 남기고 사이트를 더 적은 수의 노드로 제거할 수 있습니다.

Oracle 12cR2 이전에는 사이트 손실 중에 퇴거가 발생하는 측을 제어할 수 없었습니다. 각 사이트에 동일한 수의 노드가 있는 경우 일반적으로 부팅되는 첫 번째 RAC 노드가 마스터 노드에 의해 제거됩니다.

Oracle 12cR2에는 노드 가중치 기능이 도입되었습니다. 이 기능을 통해 관리자는 Oracle이 브레인 분할 조건을 해결하는 방법을 보다 효과적으로 제어할 수 있습니다. 간단한 예로, 다음 명령을 실행하면 RAC의 특정 노드에 대한 기본 설정이 설정됩니다.

```
[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.
```

Oracle High-Availability Services를 다시 시작한 후 구성은 다음과 같습니다.

```
[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
```

노드 `host-a` 이(가) 중요 서버로 지정되었습니다. 2개의 RAC 노드가 격리된 경우 `host-a` 존속, 그리고 `host-b` 퇴거시키다.

자세한 내용은 Oracle 백서 "Oracle Clusterware 12c Release 2 기술 개요"를 참조하십시오. ”

12cR2 이전 버전의 Oracle RAC의 경우 다음과 같이 CRS 로그를 확인하여 마스터 노드를 식별할 수 있습니다.

```
[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
```

이 로그는 마스터 노드가 임을 나타냅니다 `2` 및 노드입니다 `host-a` 의 ID가 있습니다 `1`. 이는 실제로 그 점을 의미합니다 `host-a` 은(는) 마스터 노드가 아닙니다. 마스터 노드의 ID는 명령을 사용하여 확인할 수 있습니다 `olsnodes -n`.

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

ID가 인 노드입니다 `2` 있습니다 `host-b`, 마스터 노드입니다. 각 사이트의 노드 수가 동일한 구성에서 사이트는 을(를) 사용합니다 `host-b` 어떤 이유로든 두 세트의 네트워크 연결이 끊길 경우 존속되는 사이트입니다.

마스터 노드를 식별하는 로그 항목이 시스템에서 제외될 수 있습니다. 이 경우 OCR(Oracle Cluster Registry) 백업의 타임스탬프를 사용할 수 있습니다.

```
[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
```

이 예는 마스터 노드가 임을 보여 줍니다 `host-b`. 또한 에서 마스터 노드가 변경되었음을 나타냅니다 `host-a` 를 선택합니다 `host-b` 5월 4일 2시 5분에서 21시 39분 사이. 이 마스터 노드를 식별하는 방법은 이전 OCR 백업 이후 마스터 노드가 변경될 수 있기 때문에 CRS 로그도 확인한 경우에만 사용하는 것이 안전합니다. 이 변경 사항이 발생한 경우 OCR 로그에 표시됩니다.

대부분의 고객은 전체 환경과 각 사이트에서 동일한 수의 RAC 노드를 서비스하는 단일 보팅 디스크 그룹을 선택합니다. 디스크 그룹은 데이터베이스가 포함된 사이트에 배치해야 합니다. 그 결과, 연결이 끊어지면 원격 사이트에서 제거됩니다. 원격 사이트에는 더 이상 쿼럼이 없고 데이터베이스 파일에 액세스할 수 없지만 로컬 사이트는 평소와 같이 계속 실행됩니다. 연결이 복원되면 원격 인스턴스를 다시 온라인 상태로 만들 수 있습니다.

재해가 발생할 경우 데이터베이스 파일과 보팅 디스크 그룹을 정상 사이트에서 온라인으로 전환하기 위해 전환을 수행해야 합니다. AUSO가 재해에 의해 전환을 트리거할 경우 클러스터가 동기화하고 스토리지 리소스가 정상적으로 온라인 상태가 되기 때문에 NVFAIL이 트리거되지 않습니다. AUSO는 매우 빠른 작동이며, 이전에 완료되어야 합니다 `disktimeout` 기간이 만료됩니다.

사이트는 두 곳밖에 없기 때문에 자동화된 외부 티브레이킹 소프트웨어를 사용할 수 없으며, 이는 강제 전환이 수동 작업이어야 한다는 것을 의미합니다.

3개 사이트 구성

확장된 RAC 클러스터는 3개의 사이트로 훨씬 더 쉽게 설계할 수 있습니다. MetroCluster 시스템의 절반을 호스팅하는 두 사이트도 데이터베이스 워크로드를 지원하고, 세 번째 사이트는 데이터베이스와 MetroCluster 시스템을 위한 Tiebreaker 역할을 합니다. Oracle Tiebreaker 구성은 세 번째 사이트에 투표하는 데 사용되는 ASM 디스크 그룹의 구성원을 배치하는 것만큼 간단할 수 있으며, RAC 클러스터에 홀수 노드 수가 있는지 확인하기 위해 세 번째 사이트에 운영 인스턴스를 포함할 수도 있습니다.

> 확장 RAC 구성에서 NFS를 사용하는 방법에 대한 중요한 정보는 "쿼럼 장애 그룹"에 관한 Oracle 설명서를 참조하십시오. 요약하면, 쿼럼 리소스를 호스팅하는 세 번째 사이트에 대한 연결이 끊겨 기본 Oracle 서버 또는 Oracle RAC 프로세스가 중단되지 않도록 소프트 옵션을 포함하도록 NFS 마운트 옵션을 수정해야 할 수 있습니다.