DELL Technologies

모범 사례

Dell EMC PowerFlex: 네트워킹 모범 사례 및 설계 고려 사항

PowerFlex 버전 3.5.x

개요

이 문서에서는 Dell EMC PowerFlex™ 소프트웨어 정의 스토리지의 핵심 개념, 그리고 복제를 사용한 단일 사이트 및 다중 사이트 구축을 비롯하여 PowerFlex 시스템에 대한 네트워크 설계, 문제 해결 및 유지 보수 모범 사례에 대해 설명합니다.

2021년 4월

H18390.3

개정

날짜	설명
2021년 4월	가상 네트워크 및 동적 라우팅에 대한 업데이트
2021년 1월	포함된 언어에 대한 면책 조항 추가
2020년 6월	PowerFlex 3.5 릴리스 및 브랜드 변경 – 복제 관련 내용 재작성 및 업데이트
2019년 5월	VxFlex OS 3.0 릴리스 – 추가 및 업데이트
2018년 7월	VxFlex OS 브랜드 변경 및 대대적인 재작성 – VXLAN 추가
2016년 6월	LAG 지원 범위 추가
2015년 11월	초판

감사의 말

콘텐츠 소유자: Brian Dean, 스토리지 기술 마케팅

지원: Neil Gerren, Igal Moshkovich, Matt Hobbs, Dan Aharoni, Rivka Matosevich

목차

핵심 요약

Dell EMC™ PowerFlex™ 제품군은 PowerFlex 소프트웨어 정의 스토리지에 기반하여 규모에 따라 예측 가능한 고성능 및 회복탄력성과 유연성, 탄력성, 편의성을 제공하도록 설계된 스케일 아웃 블록 스토리지 서비스입니다. 이전에 VxFlex OS라고 알려진 PowerFlex 스토리지 소프트웨어는 여러 OS 및 하이퍼바이저 기능을 포함한 여러 구축 옵션을 지원합니다.

PowerFlex 제품군은 현재 1개의 랙 수준과 2개의 노드 수준(어플라이언스 노드 및 Ready Nodes) 오퍼링으로 구성됩니다. 이 문서는 주로 스토리지 가상화 소프트웨어 계층 자체에 초점을 맞추고 있으며, 대부분 Ready Nodes와 관련된 내용이지만 성공적인 PowerFlex 기반 스토리지 시스템에 필요한 네트워킹을 이해하고자 하는 이들에게 유용한 정보를 제공합니다.

PowerFlex 랙은 모던 데이터 센터를 위해 완벽하게 엔지니어링된 랙 스케일 시스템입니다. 랙 솔루션에서 네트워킹은 사전 구성 및 최적화되며, PFxM(PowerFlex Manager)이 설계를 규정하고, 구현하고, 유지 관리합니다. 이 문서에서는 랙을 구축하는 상황을 다루지 않습니다. 다른 PowerFlex 제품군 솔루션의 경우, 반드시 하나는 적절한 네트워크를 설계하고 구현해야 합니다. PFxM 3.6 릴리스부터 어플라이언스는 특정 기준을 충족하고 PFxM이 구축한 토폴로지와 일치하도록 구성되어 있는 경우에 한하여 지원되지 않는 상용 등급 스위치를 사용할 수 있습니다. 이에 대해서는 아래에서 다룹니다.

PowerFlex 구축의 성공 여부는 설계된 네트워크 토폴로지의 적절성에 따라 달라집니다. 이 문서에서는 네트워크 옵션, 그리고 해당 옵션이 다른 PowerFlex 구성 요소 간의 트래픽 유형과 어떤 관련이 있는지에 대한 지침을 제공합니다. 이 문서에서는 하이퍼 컨버지드 고려 사항 및 소프트웨어 버전 3.5에 도입된 PowerFlex 기본 비동기식 복제를 사용한 구축과 같은 다양한 시나리오를 다룹니다. 또한 일반적인 이더넷 고려 사항, 네트워크 성능, 동적 IP 라우팅, 네트워크 가상화, VMware® 환경 내 구현, 검증 방법 및 모니터링 권장 사항을 다룹니다.

대상 및 사용

이 문서는 IT 관리자, 스토리지 설계자, Dell Technologies™ 파트너 및 직원을 대상으로 합니다. 따라서 네트워킹 전문가가 아닌 독자도 이해할 수 있습니다. 단, IP 네트워킹에 대한 중급 수준의 지식이 있다고 가정합니다.

PowerFlex(VxFlex OS)에 익숙한 독자는 "PowerFlex 기능 개요" 및 "PowerFlex 소프트웨어 구성 요소" 섹션의 대부분을 건너뛰어도 됩니다. 그러나 새로운 SDR(Storage Data Replicator) 구성 요소에 주의를 기울여야 합니다.

이 가이드에서는 네트워크 모범 사례의 일부만을 제공하며 PowerFlex의 모든 네트워킹 모범 사례 또는 구성을 다루지 않습니다. PowerFlex 기술 전문가는 이 가이드에 수록된 내용보다 더 포괄적인 모범 사례를 추천할 수도 있습니다.

Cisco Nexus® 스위치가 이 문서의 예에서 자주 사용되지만, 일반적으로 모든 네트워크 공급업체에 동일한 원칙이 적용됩니다.[1] 편의를 위해, 소비 옵션의 구분 없이 PowerFlex 노드로 하나 이상의 PowerFlex 소프트웨어 구성 요소를 실행하는 서버를 통칭합니다.

굵은 글씨로 표시되는 특정 권장 사항은 이 문서의 끝부분에 있는 "권장 사항 요약" 섹션에서 다시 살펴봅니다.

[1] Dell 네트워크 장비 사용에 대한 지침은 VxFlex Network Deployment Guide using Dell EMC Networking 25GbE switches and OS10EE 문서를 참조하십시오.

1 PowerFlex 기능 개요

PowerFlex는 직접 연결 스토리지에서 서버 및 IP 기반 SAN을 생성해 필요에 따라 성능과 용량을 유연하게 확장하는 스토리지 가상화 소프트웨어입니다. 기존 SAN 인프라의 대안인 PowerFlex는 여러 가지 성능 및 데이터 서비스 옵션을 지원하여 다양한 스토리지 미디어를 결합해 블록 스토리지로 구성된 가상 풀을 생성할 수 있습니다. PowerFlex는 엔터프라이즈급 데이터 보호, 멀티 테넌트 기능과 함께 인라인 압축, QoS, 씬 프로비저닝, 스냅샷, 기본 비동기식 복제와 같은 엔터프라이즈 기능도 제공합니다. PowerFlex의 주요 이점은 다음과 같습니다.

높은 확장성 – PowerFlex는 소수의 노드로만 시작한 후 클러스터 내에 수백 개까지 확장할 수 있습니다. 디바이스 또는 노드가 추가되면 PowerFlex가 자동으로 균등하게 데이터를 재배포함으로써 분산된 스토리지 풀의 균형을 완벽하게 유지합니다.

탁월한 성능 – PowerFlex 스토리지 풀의 모든 스토리지 미디어 디바이스는 I/O 작업을 처리하는 데 사용됩니다. 리소스에 대한 I/O 병렬 처리 덕분에 병목 현상이 사라집니다. 처리량 및 IOPS는 스토리지 풀에 추가된 스토리지 디바이스 수에 비례하여 향상됩니다. 성능 및 데이터 보호 최적화가 자동으로 이루어집니다.

뛰어난 경제성 – PowerFlex는 파이버 채널 패브릭 또는 전용 구성 요소(예: HBA)를 필요로 하지 않습니다. 구형 하드웨어에 대한 전면적인 업그레이드가 필요하지 않습니다. 고장나거나 오래된 구성 요소는 시스템에서 간단하게 제거되는 동시에 새 구성 요소가 추가되고 데이터가 재조정됩니다. PowerFlex는 이러한 방식으로 기존 SAN보다 스토리지 솔루션의 비용과 복잡성을 줄일 수 있습니다.

탁월한 유연성 – PowerFlex는 유연한 구축 옵션을 제공합니다. 2계층 구축 환경에서는 애플리케이션과 스토리지 소프트웨어가 별도의 서버 풀에 설치됩니다. 2계층 구축을 통해 컴퓨팅 팀 및 스토리지 팀은 운영의 자율성을 유지할 수 있습니다. 하이퍼 컨버지드 구축 환경에서는 애플리케이션과 스토리지가 하나의 공유 서버 풀에 설치되므로 공간을 적게 차지하고 비용이 저렴합니다. 컴퓨팅 및 스토리지 리소스를 확장할 때 이러한 구축 모델을 혼합해 사용하여 뛰어난 유연성을 제공할 수도 있습니다.

최고 수준의 탄력성 - 요구 사항이 증가할 때마다 스토리지 및 컴퓨팅 리소스를 늘리거나 줄일 수 있습니다. 필요 시 시스템에서 즉각 데이터를 자동으로 재조정합니다. 소량 또는 대량으로 추가하고 제거할 수 있으며, 용량 계획 또는 복잡한 재구성이 필요하지 않습니다. 예기치 않은 구성 요소 손실이 발생하면 데이터 보호를 유지하기 위해 재구축 작업이 트리거됩니다. 구성 요소를 추가하면 사용 가능한 성능과 용량을 높이기 위해 재조정 작업이 트리거됩니다. 재구축 및 재조정 작업은 운영자의 개입 없이 백그라운드에서 자동으로 이루어지며, 애플리케이션과 사용자에게는 다운타임이 없습니다.

엔터프라이즈 및 서비스 공급업체를 위한 필수 기능 – QoS(Quality of Service) 관리를 통해 리소스 사용량을 동적으로 관리할 수 있으며, 선택한 클라이언트가 소비할 수 있는 성능(IOPS 또는 대역폭)의 양을 제한할 수 있습니다. PowerFlex는 데이터 백업과 클론 생성을 위한 즉각적인 쓰기 가능 스냅샷을 제공합니다. 작업자는 두 가지 데이터 레이아웃 중 하나로 풀을 생성하여 워크로드에 매우 적합한 환경을 보장할 수 있습니다. 또한 요구 사항 변경이 있는 경우 서로 다른 풀 간에 실시간으로 운영 중단 없이 볼륨을 마이그레이션할 수 있습니다. 씬 프로비저닝 및 인라인 데이터 압축을 통해 스토리지를 절약하고 효율적으로 용량을 관리할 수 있습니다. 그리고 PowerFlex 버전 3.5는 재해 복구, 데이터 마이그레이션, 테스트 시나리오 및 워크로드 오프로드에 사용할 수 있는 기본 비동기식 복제를 제공합니다.

PowerFlex는 보호 도메인과 스토리지 풀을 통해 멀티 테넌트 기능을 제공하고 보호 도메인을 통해 사용자는 특정 노드 및 데이터 세트를 격리할 수 있습니다. 스토리지 풀을 사용하면 추가 데이터 분리, 계층화, 성능 관리가 가능합니다. 예를 들어 성능이 필요한 비즈니스 크리티컬 애플리케이션 및 데이터베이스용 데이터는 가장 짧은 레이턴시로 고성능 SSD, NVMe 또는 SCM 기반 스토리지 풀에 저장하고, 액세스 빈도가 낮은 데이터는 일일 쓰기 사양이 낮은 저비용의 고용량 SSD로 구축된 풀에 저장할 수 있습니다. 또한, 워크로드에 영향 없이 볼륨을 한 곳에서 다른 곳으로 실시간 마이그레이션할 수 있습니다.

2 PowerFlex 소프트웨어 구성 요소

PowerFlex는 기본적으로 SDS(Storage Data Server), SDC(Storage Data Client), MDM(Meta Data Manager)이라는 세 가지 종류의 소프트웨어 구성 요소로 구성되어 있습니다. 버전 3.5에는 복제를 지원하는 새로운 구성 요소인 SDR(Storage Data Replicator)이 도입되었습니다.

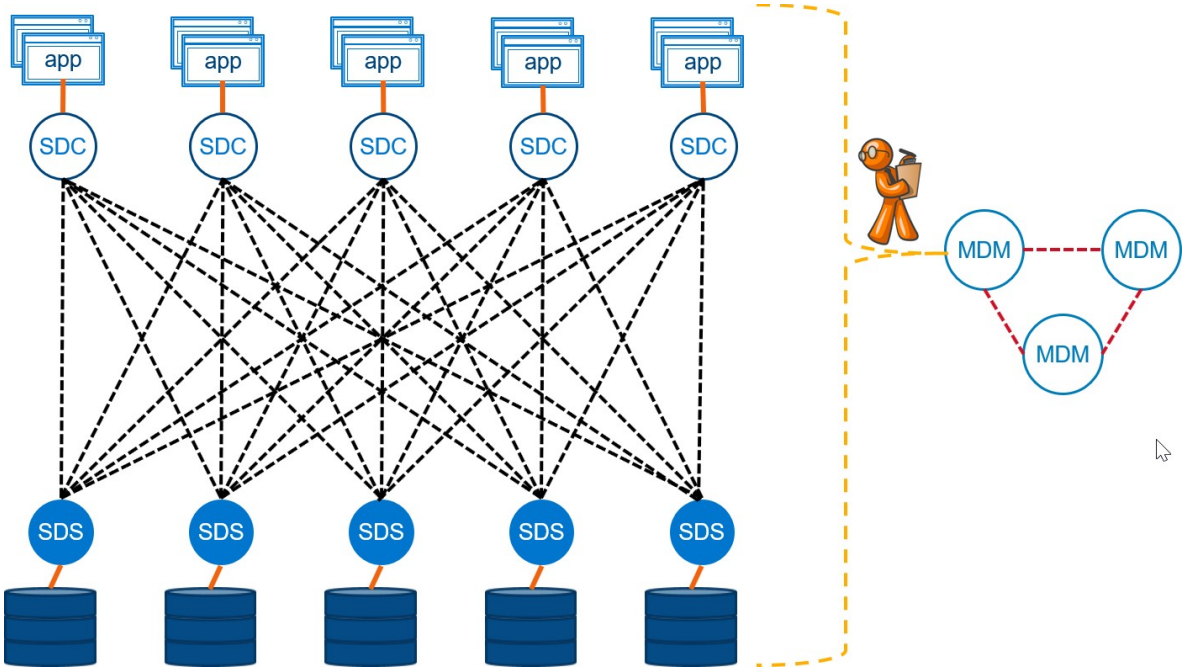

그림 1 PowerFlex 구축에 대한 논리를 설명하는 그림. SDC에서 사용할 수 있는 각 볼륨은 SDS를 실행하는 여러 시스템으로 분산되며 각 SDC는 볼륨을 제공하는 모든 SDS로의 이중화 경로를 제공합니다. MDM(Meta Data Manager) 클러스터는 시스템을 모니터링하고, 데이터 레이아웃을 조정하며, 변경 사항이 발생할 경우 SDC를 업데이트하는 데이터 경로의 외부에 상주합니다.

2.1 SDS(Storage Data Server)

SDS(Storage Data Server)는 노드의 원시 로컬 스토리지를 합쳐 PowerFlex 클러스터의 일부로 제공하는 사용자 공간 서비스입니다. SDS는 서버 측 소프트웨어 구성 요소입니다. 다른 노드로 데이터를 제공하는 모든 서버에는 해당 서버에 설치되어 실행되는 SDS 서비스가 있습니다. SDS 모음은 PowerFlex 영구 레이어를 형성합니다.

다수의 SDS가 함께 작동하여 사용자 데이터의 중복 복제본을 유지하고, 하드웨어 손실로부터 서로를 보호하며, 하드웨어 구성 요소에 장애가 발생하면 데이터 보호 기능을 재구성합니다. SDS는 SSD, PCIe 기반 플래시, 스토리지 클래스 메모리, 회전식 디스크 미디어, 가용 RAM 또는 그와 관련된 모든 조합을 활용할 수 있습니다.

SDS는 다양한 Linux 버전 또는 ESXi 기반 가상 어플라이언스에서 네이티브 방식으로 실행될 수 있습니다. PowerFlex 클러스터는 최대 512개의 SDS를 지원합니다.

SDS 구성 요소는 서로 직접 통신할 수 있으며 SDS 모음은 완전한 메시 형태로 구성됩니다. SDS는 재구축, 재조정 및 I/O 병렬 처리에 최적화되어 있습니다. SDS 구성 요소 간의 사용자 데이터 레이아웃은 **스토리지 풀, 보호 도메인 및 장애 세트**를 통해 관리됩니다.

SDC에서 사용하는 클라이언트 볼륨은 **스토리지 풀** 내부에 위치합니다. 스토리지 풀은 유사한 유형의 스토리지 미디어를 드라이브 수준으로 세분화하여 논리적으로 합치는 데 사용됩니다. 스토리지 풀은 용량과 성능에 따라 다양한 수준의 스토리지 서비스를 제공합니다.

노드, 디바이스 및 네트워크 연결 오류로부터 보호하는 기능은
보호 도메인을 통해 노드 수준으로 세분화하여 관리됩니다. 보호 도메인은 사용자 데이터 복제본이 유지되는 SDS 그룹입니다.

장애 세트를 사용하면 매우 큰 시스템에서 중복 복제본이 동시에 실패할 가능성이 높은 노드 세트(예: 전체 랙)에 상주하지 않도록 함으로써 동시에 여러 노드에서 발생하는 장애를 감당할 수 있습니다.

2.2 SDC(Storage Data Client)

SDC(Storage Data Client)를 사용하여 운영 체제 또는 하이퍼바이저가 PowerFlex 클러스터에서 제공하는 데이터에 액세스할 수 있습니다. SDC는 Windows®, 다양한 Linux 버전, IBM AIX®, ESXi® 등에서 네이티브 방식으로 실행할 수 있는 클라이언트 측 소프트웨어 구성 요소입니다. 이는 소프트웨어 HBA와 유사하지만 여러 네트워크 경로 및 엔드포인트를 병렬로 사용하는 데 최적화되어 있습니다.

SDC는 운영 체제 또는 "볼륨"이라고 하는 논리적 블록 디바이스에 액세스하여 운영 체제를 실행하는 하이퍼바이저를 제공합니다. 볼륨은 기존 SAN의 LUN과 유사합니다. 각 논리적 블록 디바이스는 데이터베이스 또는 파일 시스템용 물리적 스토리지를 제공하며, 클라이언트 노드에 로컬 디바이스로 표시됩니다.

SDC는 볼륨의 블록 위치를 기준으로 연결할 SDS(Storage Data Server) 엔드포인트를 식별합니다. SDC는 PowerFlex를 실행하는 다른 시스템에서 직접 분산된 스토리지 리소스를 사용합니다. SDC는 단일 프로토콜 타겟 또는 네트워크 엔드포인트를 다른 SDC와 공유하지 않습니다. SDC는 부하를 균등하게 자율적으로 배포합니다.

SDC는 매우 가볍습니다. SDC에서 SDS로의 통신은 본질적으로 스토리지 풀에 기여하는 모든 SDS 스토리지 서버에서 다중 경로를 통해 이루어집니다. 여러 클라이언트가 단일 프로토콜 엔드포인트를 대상으로 하는 iSCSI와 같은 접근 방식과 상반됩니다. 광범위하게 분산시키는 SDC 통신의 특성은 훨씬 더 뛰어난 성능과 확장성을 제공합니다.

SDC를 사용하면 클러스터링과 같은 용도로 볼륨 액세스 권한을 공유할 수 있습니다. SDC에는 iSCSI 이니시에이터, 파이버 채널 이니시에이터 또는 FCoE 이니시에이터가 필요하지 않습니다. SDC는 편의성, 속도 및 효율성에 최적화되어 있습니다. PowerFlex 클러스터는 최대 1024개의 SDC를 지원합니다.

2.3 MDM(Meta Data Manager)

MDM은 PowerFlex 시스템의 동작을 제어합니다. MDM은 클라이언트와 해당 볼륨 데이터 간의 매핑을 결정 및 게시하고, 시스템의 상태를 추적하고, SDS 구성 요소에 대한 재구축 및 재조정 지침을 배포합니다.

MDM은 PowerFlex에서 쿼럼의 개념을 정립하며 PowerFlex에서 긴밀하게 클러스터링된 유일한 구성 요소입니다. MDM은 신뢰할 수 있고, 이중화를 지원하며, 가용성이 높습니다. 이는 I/O 작업 중 또는 재구축 및 재조정 등의 SDS 간 작업 중에 참조되지 않습니다. 하드웨어 구성 요소에 장애가 발생하더라도 MDM 클러스터는 자동 복구 작업을 몇 초 내에 시작하도록 지시합니다. MDM 클러스터는 쿼럼을 유지하기 위해 최소 3대의 서버로 구성되지만 5대를 사용하여 가용성을 개선할 수 있습니다. 3노드 또는 5노드 MDM 클러스터에는 항상 1개의 기본 노드가 있습니다. 1~2개의 보조 MDM과 1~2개의 Tie Breaker가 있을 수 있습니다.

2.4 SDR(Storage Data Replicator)

버전 3.5부터 PowerFlex 클러스터 간 비동기식 복제를 지원하는 새로운 소프트웨어 구성 요소(선택 사항)가 도입되었습니다. 복제를 사용하지 않는 경우 일반적인 PowerFlex 작업에는 SDR(Storage Data Replicator)이 필요하지 않습니다. 소스 측에서 SDR은 SDC, 그리고 볼륨 주소 공간의 관련 부분을 호스팅하는 SDS 사이의 중개자 역할을 합니다. SDC는 볼륨이 복제되면 SDR에 쓰기를 보냅니다. 이때 쓰기는 분할된 후 분할분 모두 복제 저널에 기록되고, 관련 SDS 서비스로 전달되어 로컬 디스크로 커밋됩니다.

SDR은 MDM이 간격을 종료하라고 지시할 때까지 일정 간격으로 저널에 쓰기를 누적합니다. 볼륨이 다중 볼륨 복제 정합성 보장 그룹에 속하는 경우 간격 종료가 동시에 이루어집니다. 쓰기 폴딩이 적용되며, 타겟 측으로 전송하기 위한 전송 큐에 간격이 추가됩니다.

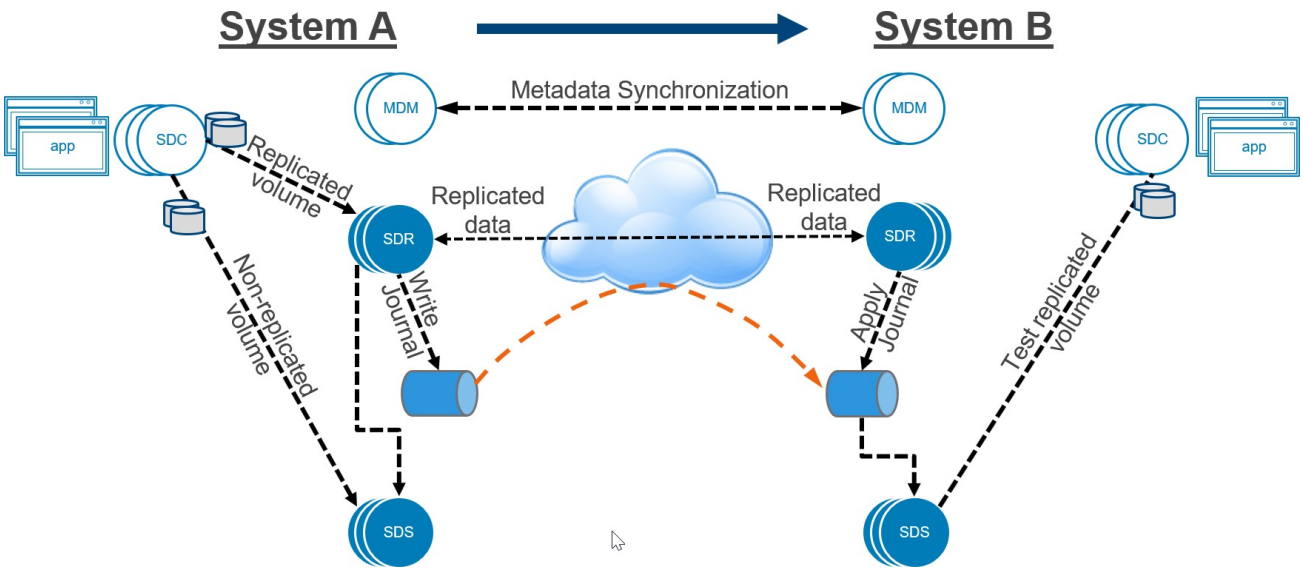

그림 2 복제 데이터 흐름을 간소화한 다이어그램.

타겟 측에서 SDR은 다른 저널에 데이터를 수신하고 이를 타겟 복제본 볼륨의 애플리케이션에서 사용하는 SDS로 보냅니다.

3 트래픽 유형

네트워크 아키텍처에 PowerFlex 트래픽 패턴을 반영하는 경우, PowerFlex의 성능, 확장성 및 보안 기능을 통해 이점을 누릴 수 있습니다. 대규모로 PowerFlex를 구축하는 경우에 특히 그렇습니다. PowerFlex를 구성하는 소프트웨어 구성 요소(SDC, SDS, MDM, SDR)는 예측 가능한 방식으로 서로 통신합니다. **PowerFlex 구축을 설계하는 설계자는 이러한 트래픽 패턴을 알아야 네트워크 레이아웃에 대해 정보에 입각한 선택을 내릴 수 있습니다.**

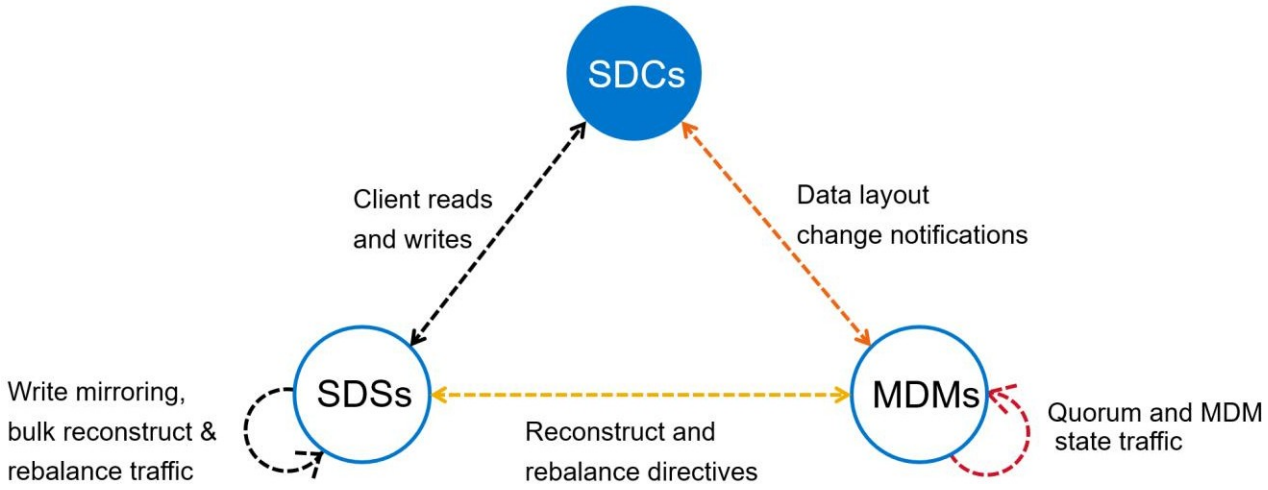

그림 3 기본 PowerFlex 소프트웨어 구성 요소가 통신하는 방법에 대한 간단한 그림. PowerFlex 시스템에는 많은 SDC, SDS 및 MDM이 있습니다. 이 그림에서는 SDC, SDS 및 MDM을 그룹화합니다. SDS 및 MDM에서 자신을 가리키는 화살표는 다른 SDS 및 MDM과의 통신을 나타냅니다. SDC 간의 통신은 없습니다. 이 트래픽 패턴은 SDC, SDS 또는 MDM의 물리적 위치와 관계없이 동일합니다.

다음 논의에서는 백엔드 트래픽과 프런트엔드 트래픽을 구분합니다. 이는 논리적 구분이며 물리적으로 별개의 네트워크가 필요하지 않습니다. PowerFlex는 동일한 물리적 네트워크를 통해 프런트엔드 및 백엔드 트래픽을 모두 실행하거나 이를 별개의 네트워크로 분리할 수 있습니다. 필수는 아니지만 스토리지 네트워크의 프런트엔드 및 백엔드 트래픽을 격리시키는 것이 좋습니다.

예를 들어, 운영상의 이유로 이렇게 분리할 수 있으며, 이 경우에는 별도의 팀이 인프라스트럭처의 서로 다른 부분을 관리합니다. 그러나 백엔드 트래픽을 분리하는 가장 일반적인 이유는 재구축 및 재조정 성능을 높일 수 있기 때문입니다. 또한 프런트엔드 트래픽을 격리시킴으로써 네트워크에서의 경합을 방지하고 재구축/재조정 작업 중에 클라이언트 또는 애플리케이션 트래픽에 레이턴시가 미치는 영향을 줄일 수 있습니다.

3.1 SDC(Storage Data Client)에서 SDS(Storage Data Server)로

SDC와 SDS 간의 트래픽은 대량의 프런트엔드 스토리지 트래픽을 형성합니다. 프런트엔드 스토리지 트래픽에는 클라이언트에 도착하거나 클라이언트에서 발생한 모든 읽기 및 쓰기 트래픽이 포함되며 이 네트워크에는 높은 처리량이 필요합니다.

3.2 SDS(Storage Data Server)에서 SDS(Storage Data Server)로

SDS 간의 트래픽은 대량의 백엔드 스토리지 트래픽을 형성합니다. 백엔드 스토리지 트래픽에는 SDS, 재조정 트래픽, 재구축 트래픽, 볼륨 마이그레이션 트래픽 사이에 미러링되는 쓰기가 포함되며 이 네트워크에는 높은 처리량이 필요합니다.

3.3 MDM(Meta Data Manager)에서 MDM(Meta Data Manager)으로

MDM은 클러스터 내부 작업을 조정하는 데 사용되며 PowerFlex로 트래픽을 재조정, 재구축 및 리디렉션하도록 지침을 배포합니다. 또한 복제 정합성 보장 그룹을 조정하고, 복제 저널 간격 종료를 결정하고, PowerFlex 복제본 피어 시스템과의 메타데이터 동기화를 유지합니다. MDM은 이중화되고 지속적으로 서로 통신해야 쿼럼을 정립하고 데이터 레이아웃에 대한 정보를 계속 공유할 수 있습니다.

MDM은 I/O 트래픽을 전달하거나 I/O 트래픽에 직접 간섭하지 않습니다. 그중 교환된 데이터는 상대적으로 가벼우며 MDM은 SDS 또는 SDC 트래픽에 필요한 것과 동일한 수준의 처리량을 필요로 하지 않습니다. 단, MDM의 경우 100ms마다 발생하는 쿼럼 교환의 시간 제한이 매우 짧습니다(400ms 미만).
MDM 간의 트래픽에는 안정적이며 신뢰성이 높고 레이턴시가 짧은 네트워크가 필요합니다. MDM 간 트래픽은 백엔드 스토리지 트래픽으로 간주됩니다. PowerFlex는 MDM 간의 트래픽에 하나 이상의 전용 네트워크를 사용할 수 있도록 지원합니다. 운영 환경에서는 MDM당 최소 2개의 10GbE 링크를 사용해야 하지만, 25GbE가 더 일반적입니다.

PowerFlex 3.5은 복제 피어 시스템 간 MDM 트래픽에 클러스터 간 MDM을 지원합니다. 이러한 MDM은 통신을 통해 복제 흐름과 저널 상태를 제어해야 하며 소스 및 대상 사이트 간의 통합 복제 상태를 동기화합니다. MDM 간 피어 메타데이터 동기화는 200ms 미만의 레이턴시로 WAN을 통해 이루어져야 합니다.

3.4 MDM(Meta Data Manager)에서 SDC(Storage Data Client)로

데이터 레이아웃이 변경되는 경우 기본(소프트웨어가 마스터를 호출하는 것) MDM은 SDC와 통신해야 합니다. 이러한 상황은 SDS를 위해 SDC의 볼륨 스토리지를 호스팅하는 SDS가 추가되거나, 제거되거나, 유지 보수 모드에 있거나, 오프라인 상태로 전환됨으로 인해 발생할 수 있습니다. 볼륨을 복제 정합성 보장 그룹에 배치한 경우에도 발생할 수 있습니다. 기본 MDM과 SDC 간의 통신은 느리고 비동기식 지만 여전히 신뢰성이 높고 레이턴시가 짧은 네트워크를 필요로 합니다. MDM에서 SDC로의 트래픽은 프런트엔드 스토리지 트래픽으로 간주됩니다.

3.5 MDM(Meta Data Manager)에서 SDS(Storage Data Server)로

기본 MDM은 SDS와 통신하여 SDS 및 디바이스 상태를 모니터링하고 재조정 및 재구축 지침을 배포해야 합니다. MDM에서 SDS로의 트래픽은 신뢰성이 높고 레이턴시가 짧은 네트워크를 필요로 합니다. MDM에서 SDS로의 트래픽은 백엔드 스토리지 트래픽으로 간주됩니다.

3.6 SDC(Storage Data Client)에서 SDR(Storage Data Replicator)로

볼륨이 복제되는 경우 일반 SDC에서 SDS로의 트래픽은 SDR을 통해 라우팅됩니다. 볼륨이 복제 정합성 보장 그룹에 배치되는 경우 MDM은 SDC에 부여되는 볼륨 매핑을 조정하고, I/O 작업을 SDR로 배포하라고 SDC에 지시합니다. 그러면 SDR은 이를 관련 SDS로 전달합니다. SDR은 또 다른 SDS인 것처럼 SDC에 나타납니다. SDC에서 SDR로의 트래픽은 높은 처리량과 신뢰성 높고 레이턴시가 짧은 네트워크를 필요로 합니다. SDC에서 SDR로의 트래픽은 프런트엔드 스토리지 트래픽으로 간주됩니다.

3.7 SDR(Storage Data Replicator)에서 SDS(Storage Data Server)로

볼륨이 복제되고 I/O가 SDC에서 SDR로 전송되면 SDR에서 SDS로의 두 가지 후속 I/O가 소스 시스템에서 이루어집니다. 먼저 SDR이 볼륨 I/O를 연결된 SDS로 전달하여 처리(예: 압축)하고 디스크에 커밋합니다. 두 번째는 SDR이 저널링 볼륨에 쓰기를 적용합니다. 저널 볼륨은 PowerFlex 시스템의 또 다른 볼륨에 불과하므로 SDR은 디스크가 저널 볼륨이 상주하는 스토리지 풀로 구성된 SDS로 I/O를 전송합니다.

타겟 시스템에서 SDR은 수신된 동일한 저널을 복제본 볼륨의 기반이 되는 SDS에 적용합니다. 이러한 경우에 SDR은 SDC처럼 작동합니다. 그럼에도 불구하고 SDR에서 SDS로의 트래픽은 백엔드 스토리지 트래픽으로 간주됩니다. SDR에서 SDS로의 트래픽 처리량은 매우 높을 수 있으며 복제되는 볼륨 수에 비례합니다. 이는 신뢰성이 높고 레이턴시가 짧은 네트워크를 필요로 합니다.

3.8 MDM(Meta Data Manager)에서 SDR(Storage Data Replicator)로

MDM은 저널 간격 종료를 발행하고, RPO 규정 준수를 수집 및 보고하며, 대상 볼륨에서 일관성을 유지하기 위해 SDR과 통신해야 합니다. MDM은 피어 시스템에서 전송되는 복제 상태를 사용하여 로컬 SDR에 저널 작업을 수행할 것을 명령합니다.

3.9 SDR(Storage Data Replicator)에서 SDR(Storage Data Replicator)로

소스 내 또는 타겟 PowerFlex 클러스터 내에 있는 SDR은 서로 통신하지 않습니다. 하지만 소스 시스템의 SDR은 복제본 타겟 시스템에서 SDR과 통신합니다. SDR은 LAN 또는 WAN 네트워크를 통해 저널 간격을 대상 SDR로 전송합니다. SDR → SDR 트래픽의 경우 레이턴시가 중요하지는 않지만 왕복 시간은 200ms 이하여야 합니다.

3.10 기타 트래픽

PowerFlex 클러스터에는 기타 여러 가지 유형의 저용량 트래픽이 있습니다. 기타 트래픽으로는 드물게 발생하는 관리, 설치 및 보고 트래픽이 포함됩니다. 또한 PowerFlex Gateway(REST API 게이트웨이, 설치 관리자, SNMP 트랩 발신기), vSphere 플러그인, PowerFlex Manager, LIA(Light Installation Agent)를 오가는 트래픽과 MDM으로의 보고 또는 관리 트래픽(예: 보고를 위한 syslog, 관리 인증을 위한 LDAP)이 포함됩니다. MDM, SDS, SDC 간 CHAP 인증 트래픽도 포함됩니다. 자세한 내용은 PowerFlex 기술 리소스 센터의 "Dell EMC PowerFlex 알아보기" 가이드를 참조하십시오.

SDC는 다른 SDC와 통신하지 않습니다. 이는 사설 VLAN 및 네트워크 방화벽을 사용하여 적용할 수 있습니다.

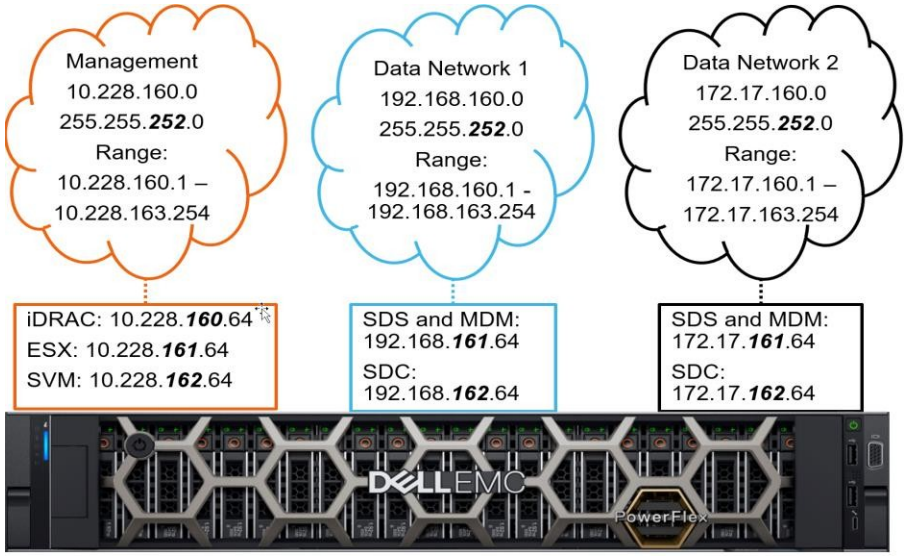

그림 4 간단한 PowerFlex 하이퍼 컨버지드 네트워크 레이아웃. 관리 네트워크가 라우팅되고 iDRAC, ESX 및 SVM(Storage Virtual Machine)에 대한 액세스를 제공합니다. 이중화된 네트워크는 SDS, MDM 및 SDC 트래픽을 전달합니다. SDS 및 MDM 트래픽은 동일한 IP 주소 세트를 사용합니다. 트래픽은 대규모 구축 환경처럼 프런트엔드 트래픽(SDS, SDC, MDM)과 백엔드 트래픽(SDS, MDM)으로 분할되지 않습니다. MDM 가상 IP에는 192.168.160.X 및 172.17.160.X 주소 공간을 사용할 수 있습니다.

4 PowerFlex TCP 포트 사용

PowerFlex는 이더넷 패브릭을 통해 작동합니다. 대부분의 PowerFlex 프로토콜은 독점적이지만 모든 통신에는 표준 TCP/IP 전송이 사용됩니다.

다음 다이어그램은 PowerFlex 소프트웨어 구성 요소 간의 포트 사용 및 통신에 대한 개요를 제공합니다. 일부 포트는 고정되어 변경되지 않을 수 있지만 다른 포트는 구성 가능하고 다른 포트로 재할당할 수 있습니다. 전체 목록 및 분류에 대한 자세한 내용은 Dell EMC PowerFlex Security Configuration Guide의 "포트 사용 및 기본 포트 변경" 섹션을 참조하십시오.

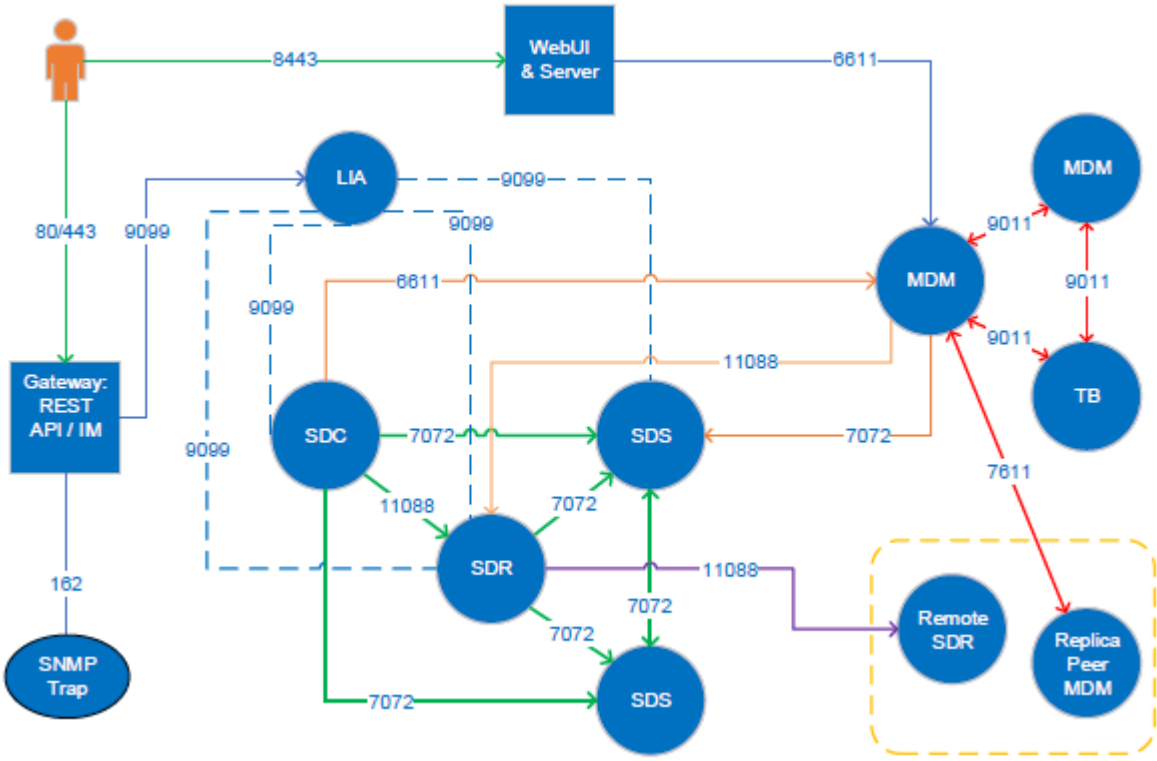

그림 5 PowerFlex 소프트웨어 정의 스토리지 구성 요소 내 TCP 포트 사용 및 통신. 다이어그램의 화살표는 연결 시작 방향을 나타냅니다. 즉, 화살표는 수신 대기 서비스 포트를 가리킵니다. 시작 후 데이터는 연결을 통해 양방향으로 이동할 수 있습니다. 파선은 설치된 구성 요소 중 노드 내부에서 통신이 이루어짐을 나타냅니다.

MDM의 포트 25620 및 25600, 그리고 SDS의 포트 25640도 수신 대기할 수 있습니다. 이러한 포트는 PowerFlex 내부 디버깅 툴에서만 사용되며 일상적인 운영 및 트래픽에는 해당하지 않습니다.

5 네트워크 내결함성

PowerFlex 구성 요소(MDM, SDS, SDC, SDR) 간의 통신은 서로 다른 물리적 네트워크에 있는 두 개 이상의 서브넷에 할당되어야 합니다. 이러한 각 구성 요소로 구성된 PowerFlex 네트워킹 레이어는 할당된 여러 서브넷에 기본 링크 내결함성 및 경로 다중화를 제공합니다. 이로 인한 설계상의 장점은 다음과 같습니다.

1. 링크 오류가 발생하면 PowerFlex는 문제를 거의 즉시 인식하고 대역폭 손실에 맞게 조정합니다.
2. 스위치 기반 링크 집선을 사용한 경우 PowerFlex는 단일 링크 손실을 식별할 수 있는 수단이 없습니다.
3. PowerFlex는 링크에 실패할 경우 MDM, SDS 및 SDC 구성 요소에 할당된 서브넷에서 2~3초 안에 통신을 동적으로 조정합니다. 이는 SDS→SDS 및 SDC→SDS 연결 시 특히 중요합니다.
4. 이러한 각 구성 요소는 최대 8개의 서브넷에서 트래픽을 로드 밸런싱하고 집계할 수 있으므로 스위치 기반 링크 집선을 유지 관리하는 데 따르는 복잡성이 줄어듭니다. 그리고 스토리지 계층 자체에서 관리되기 때문에 스위치 기반 집선보다 더 효율적이고 유지 관리가 더 간단합니다.

   참고: 이전 버전의 PowerFlex 소프트웨어에서는 링크 관련 오류가 발생하는 경우 SDC→SDS 네트워크에서 네트워크 서비스 중단과 최대 17초의 I/O 지연이 발생할 수 있습니다. SDC에는 대개 15초의 제한 시간이 있으며, 제한 시간에 도달했고 데드 소켓이 이미 닫혀 있으면 I/O가 다른 "정상" 소켓에서만 다시 이루어집니다.

   버전 3.5 이상에서 PowerFlex는 더 이상 I/O 제한 시간을 사용하지 않고 링크 연결 해제 알림을 사용합니다. 링크 중단이 발생하면 관련된 모든 TCP 연결이 2초 후에 닫히고, 응답을 받지 못한 전송 중인 모든 I/O 메시지가 중단되고 해당 I/O는 SDC에 의해 다시 이루어집니다.

기본 네트워크 경로 로드 밸런싱 및 스위치 기반 링크 집선이 모두 완벽하게 지원되지만, 기본 네트워크 경로 로드 밸런싱을 사용하는 편이 더 간단합니다. 원하는 경우, 예를 들어, 각 논리적 네트워크가 노드당 2개의 물리적 포트를 사용하는 트렁크에서 두 개의 데이터 경로 네트워크를 생성하는 방식을 함께 사용할 수 있습니다.

어플라이언스를 사용하는 경우, PowerFlex Manager는 반드시 이 작업을 수행하며 기본 경로 다중화와 함께 링크 집선을 사용하여 계층화된 강력한 네트워크 내결함성을 제공합니다. Dell EMC PowerFlex Appliance Network Planning Guide를 참조하십시오.

6 네트워크 인프라스트럭처

현재 PowerFlex에는 리프-스파인 및 플랫 네트워크 토폴로지가 가장 일반적으로 사용됩니다. 소규모 네트워크에는 플랫 네트워크가 사용됩니다. 최신 데이터 센터에서는 기존 계층형 토폴로지보다 리프-스파인 토폴로지를 선호합니다. 이 섹션에서는 PowerFlex 데이터 트래픽의 전송 매체인 플랫 토폴로지와 리프-스파인 토폴로지를 비교합니다.

Dell Technologies는 비차단 네트워크 설계를 사용할 것을 권장합니다. 비차단 네트워크 설계를 사용하면 메시지 루프를 방지하기 위해 네트워크 포트 일부를 차단하지 않고 모든 스위치 포트를 동시에 사용할 수 있습니다. 따라서 Dell Technologies는 PowerFlex를 호스팅하는 네트워크에 STP(Spanning Tree Protocol)를 사용할 것을 적극 권장합니다. 극대화된 성능과 예측 가능한 QoS(Quality of Service)를 달성하려면 네트워크가 초과 할당되어서는 안 됩니다.

6.1 리프-스파인 네트워크 토폴로지

2계층 리프-스파인 토폴로지는 리프 스위치 사이에 단일 스위치 홉을 제공하고 엔드포인트 사이에 대량의 대역폭을 제공합니다. 적절한 규모의 리프-스파인 토폴로지에서는 업링크 포트의 초과 할당이 이루어지지 않습니다. 규모가 매우 큰 데이터 센터는 3계층 리프-스파인 토폴로지를 사용할 수 있습니다. 이 백서에서는 2계층 리프-스파인 구축에 초점을 맞춥니다.

리프-스파인 토폴로지에서 각 리프 스위치는 모든 스파인 스위치로 연결됩니다. 리프 스위치는 다른 리프 스위치에 직접 연결할 필요가 없습니다. 스파인 스위치는 다른 스파인 스위치에 직접 연결할 필요가 없습니다.

Dell Technologies는 대부분의 경우에 리프-스파인 네트워크 토폴로지를 사용할 것을 권장합니다. 그 이유는 다음과 같습니다.

- PowerFlex는 단일 클러스터에서 수백 개의 노드로 스케일 아웃할 수 있습니다.
- 리프-스파인 아키텍처로 향후 환경 변화에 민첩하게 대응할 수 있습니다. 네트워크를 다시 설계할 필요 없는 간편한 스케일 아웃 구축을 지원합니다.
- 리프-스파인 토폴로지를 사용하면 모든 네트워크 링크를 동시에 사용할 수 있습니다. 기존 계층형 토폴로지는 루프를 방지하기 위해 일부 포트를 차단하는 STP(Spanning Tree Protocol)와 같은 기술을 사용해야 합니다.
- 적절한 규모의 리프-스파인 토폴로지에서는 업링크 초과 할당이 없으므로 레이턴시를 더욱 정확히 예측할 수 있습니다.

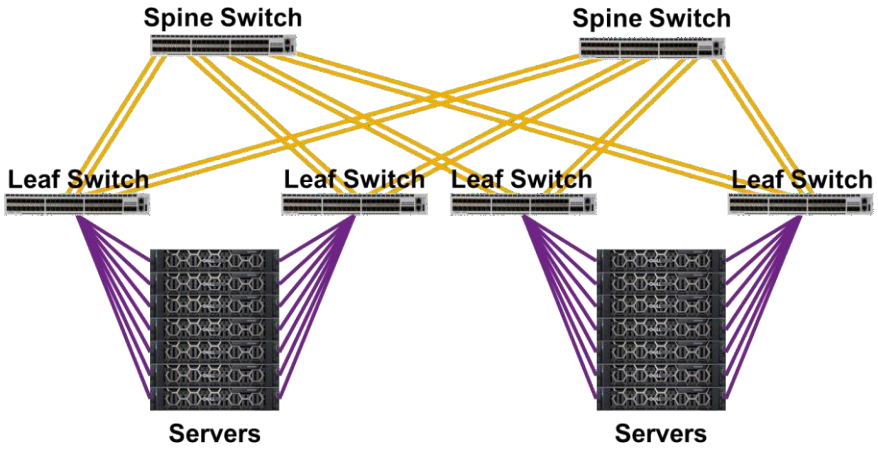

그림 6 2계층 리프-스파인 네트워크 토폴로지. 각 리프 스위치에는 다른 모든 리프 스위치로의 경로가 여러 개 있습니다. 모든 링크가 활성 상태입니다. 따라서 네트워크의 디바이스 간 처리량이 증가합니다. 리프 스위치를 서로 연결하여 MLAG에 사용할 수도 있습니다(표시되지 않음).

6.2 플랫 네트워크 토폴로지

플랫 네트워크 토폴로지는 구현하기 더 쉬울 수 있으며 기존 플랫 네트워크가 확장되고 있거나 네트워크를 확장할 것으로 예상되지 않는 경우 선호하는 옵션이 될 수도 있습니다. 플랫 네트워크에서는 모든 스위치가 호스트 연결에 사용됩니다. 스파인 스위치는 없습니다.

그러나 소수의 액세스 스위치 이상으로 확장하는 경우 추가 교차 링크 포트를 사용하면 플랫 네트워크 토폴로지 비용이 엄청나게 늘어날 수 있습니다. 플랫 네트워크 토폴로지의 활용 사례로는 규모가 랙 몇 개 수준 이상으로 커지지 않는 개념 증명 구축 및 소규모 데이터 센터 구축이 있습니다.

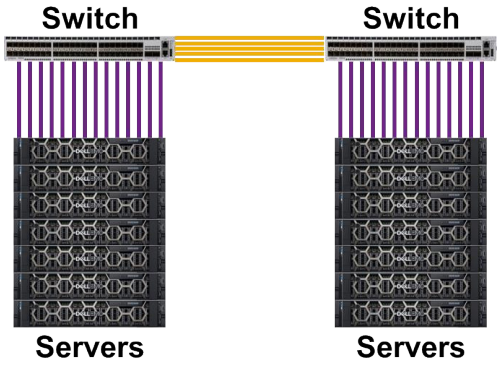

그림 7 플랫 네트워크. 이 네트워크 설계는 이중화와 확장성을 희생하는 대신 비용과 복잡성이 낮습니다. 이 그림에서 각 스위치는 단일 장애 지점입니다. MLAG(표시 되지 않음)와 같은 기술을 사용하면 단일 장애 지점 없이 플랫 네트워크를 구축할 수 있습니다.

7 네트워크 성능 및 사이징

적절한 규모의 네트워크를 사용하면 네트워크 및 스토리지 관리자가 개별 포트 또는 링크가 성능 또는 운영 병목현상을 일으킬까 염려하지 않아도 됩니다. 엔드포인트 핫 스팟 대신 네트워크를 관리하는 것이 PowerFlex의 핵심 아키텍처의 장점입니다.

PowerFlex는 네트워크의 여러 지점에 I/O를 균등하게 분배하므로 네트워크 성능을 적절히 조정해야 합니다.

7.1 네트워크 레이턴시

네트워크 레이턴시는 네트워크를 설계할 때 고려해야 하는 중요한 요소입니다. 네트워크 레이턴시를 최소화하면 성능 및 신뢰성이 향상됩니다. **최상의 성능을 얻으려면 모든 SDS 및 SDC 통신의 레이턴시가 정상 작동 조건에서 네트워크만을 왕복하는 시간이 1밀리초를 초과해서는 안 됩니다.** WAN(Wide Area Network)의 가장 낮은 응답 시간이 대개 이 제한을 초과하므로 WAN에서 PowerFlex 클러스터를 작동해서는 안 됩니다.

일반, SDC, MDM 및 SDS 통신과 관련하여 비동기식 복제를 구현하는 시스템도 예외는 아닙니다. 데이터는 독립적인 PowerFlex 클러스터 간에 복제되며, 각각은 1ms 미만이라는 이 규칙을 준수해야 합니다. 피어 시스템 간 레이턴시의 경우는 다릅니다. 비동기식 복제는 일반적으로 WAN을 통해 이루어지기 때문에 레이턴시 요구 사항은 제약이 덜합니다. **그러나 피어로 이루어진 PowerFlex 클러스터 구성 요소 간의 네트워크 레이턴시는 MDM←→MDM이든, SDR←→SDR이든, 왕복 시간이 200ms를 초과해서는 안 됩니다.**

레이턴시는 모든 구성 요소 간 양방향으로 테스트해야 합니다. 이는 Ping을 통해 검증할 수 있으며, SDS 네트워크 레이턴시 측정 테스트를 통해 더 엄격히 검증할 수도 있습니다. 오픈 소스 툴인 iPerf를 사용하면 대역폭을 검증할 수 있습니다. 단, iPerf는 Dell Technologies에서 지원하지 않습니다. PowerFlex 구축 검증에 사용된 iPerf 및 기타 툴은 이 문서의 "검증 방법" 섹션에서 자세히 다룹니다.

7.2 네트워크 처리량

네트워크 처리량은 PowerFlex 구현 설계 시 중요한 구성 요소입니다. 처리량은 장애가 발생한 노드가 재구축되는 데 걸리는 시간을 줄이고, 데이터 분포가 고르지 않은 경우 데이터를 재배포하는 데 걸리는 시간을 단축하고, 노드가 제공할 수 있는 I/O의 양을 최적화하하고, 성능 기대치를 충족하는 데 중요합니다.

PowerFlex 소프트웨어를 테스트 또는 조사를 위해 1기가비트 네트워크에 배포할 수 있지만, 네트워크 용량에 의해 스토리지 성능에 병목현상이 발생할 가능성이 높습니다. **Dell Technologies에서는 최소한도로 10기가비트 네트워크 기술을 활용할 것을 권장하나, 기본적인 최소 링크 처리량을 발휘하려면 25기가비트 기술을 사용하는 것이 좋습니다.** 현재 모든 PowerFlex 노드는 최소 4개의 포트를 제공하며, 각각에는 미래에 대비한 옵션인 100GbE 포트와 함께 최대 25GbE의 포트 대역폭이 제공됩니다. 이는 복제 사례와 추가 대역폭 요구 사항을 고려할 때 특히 중요합니다.

또한 PowerFlex 클러스터 자체는 이기종일 수 있지만, **보호 도메인을 구성하는 SDS 구성 요소는 스토리지와 네트워크 성능이 동일한 하드웨어에 상주해야 합니다**. 이는 기여하는 모든 구성 요소의 볼륨 데이터에 대한 광범위한 스트라이핑으로 인해 보호 도메인의 총 대역폭이 I/O 및 재구축/재조정 작업 중에 가장 약한 링크에 의해 제약을 받기 때문입니다. 산악회에서 가장 느린 일행보다 빨리 이동하지 못하는 상황을 떠올려 보십시오.

이기종 OS 및 하이퍼바이저 조합을 혼합할 때에도 비슷한 고려 사항이 있습니다. VMware 기반 하이퍼컨버지드 인프라스트럭처는 가상화 오버헤드로 인해 베어 메탈 구성보다 느린 성능 프로파일을 제공하며, 보호 도메인에 HCI 및 베어 메탈 노드를 혼합하면 두 노드에서 가장 느린 구성원의 성능을 포함하여 스토리지 풀의 처리량이 제한됩니다. 이러한 구성이 스토리지 소프트웨어 관점에서는 가능하고 허용되지만 사용자는 그 영향에 유의해야 합니다. PowerFlex 랙 또는 어플라이언스에서 지원되는 구성이 아닙니다.

처리량 관련 고려 사항 외에도 **처리량 요구 사항과 관계없이 이중화를 위해 각 노드를 두 개 이상의 네트워크에 별도로 연결하는 것이 좋습니다.** 네트워크 기술이 향상되더라도 이렇게 구성해야 합니다. 예를 들어, 2개의 40기가비트 링크를 단일 100기가비트 링크로 교체하면 처리량은 향상되지만 링크 수준 네트워크 이중화를 희생해야 합니다.

대부분의 경우 노드에 대한 네트워크 처리량은 노드에서 호스팅된 스토리지 미디어의 총 최대 처리량 이상이어야 합니다. *다시 말해 노드의 네트워크 요구 사항은 기본 스토리지 미디어의 총 성능에 비례합니다.*

필요한 네트워크 처리량을 결정할 때 최신 미디어 성능은 일반적으로 초당 *메가바이트*로 측정되지만 최신 네트워크 링크는 일반적으로 초당 *기가비트*로 측정됩니다.

초당 *메가바이트*를 초당 *기가비트*로 변환하려면 먼저 *메가바이트*에 8을 곱하여 *메가비트*로 변환한 후, *메가비트*를 1,000으로 나누면 *기가비트*를 구할 수 있습니다.

$$\text{기가비트} = \frac{\text{메가바이트} * 8}{1{,}000}$$

이 공식은 PowerFlex의 표준으로써 2의 제곱으로 1,024가 되는 "킬로"의 정의를 고려하지 않으므로 완벽하게 정확하지는 않지만 이 백서에서 설명하는 용도로는 충분합니다.

7.2.1 예: SSD 10개를 사용하는 SDS 전용(스토리지 전용) 노드

SDS만 호스팅하는 1U 노드가 있다고 가정합니다. 이는 하이퍼 컨버지드 환경이 아니므로 스토리지 트래픽만 고려해야 합니다. 노드에는 SAS SSD 드라이브 10개가 있습니다. 이러한 각 드라이브는 최적 조건(재구성 및 재구축 작업 중에 PowerFlex의 최적화 기준이 되는 순차 I/O)에서 초당 1,000메가바이트의 원시 처리량을 개별적으로 제공할 수 있습니다. 따라서 기본 스토리지 미디어의 총 처리량은 초당 10,000*메가바이트*입니다.

$$10 * 1{,}000\text{메가바이트} = 10{,}000\text{메가바이트}$$

그런 다음 앞에서 설명한 공식을 사용하여 먼저 10,000MB에 8을 곱한 후 1,000으로 나누어 10,000*메가바이트*를 *기가비트*로 변환합니다.

$$\frac{10{,}000\text{메가바이트} * 8}{1{,}000} = 80\text{기가비트}$$

이때 노드의 모든 드라이브가 가능한 최고 속도로 읽기 작업을 처리하는 경우 필요한 총 네트워크 처리량은 초당 80기가비트입니다. 일반적으로 네트워크 대역폭 요구 사항을 예측하기에 충분한 읽기 작업만을 고려하고 있습니다. 단일 25기가비트 링크

또는 40기가비트 링크로는 요구 사항을 충족할 수 없습니다. 이론적으로는 100GbE 링크로 충분합니다. 하지만 네트워크 이중화가 권장되므로 이 노드에는 최소 2개의 40기가비트 링크가 있어야 하며, 4개의 표준 25GbE 구성을 사용하는 것이 좋습니다.

참고: 구성 요소 드라이브의 이론적 처리량만을 기준으로 처리량을 계산하면 단일 노드에 대해 매우 높은 추정치가 나올 수 있습니다. **노드의 RAID 컨트롤러 또는 HBA가 기본 스토리지 미디어의 최대 처리량 이상을 발휘할 수 있는지 확인하십시오.**

7.2.2 쓰기 집약적 환경

읽기 및 쓰기 작업은 PowerFlex 환경에서 다양한 트래픽 패턴을 생성합니다. 호스트(SDC)는 단일 4k 읽기 요청 시 단일 SDS에 연결하여 데이터를 검색합니다. 4k 블록은 단일 SDS에서 한 번 전송됩니다. 해당 호스트가 단일 4k 쓰기 요청을 하는 경우 4k 블록을 기본 SDS로 전송한 다음 기본 SDS에서 보조 SDS로 복제해야 합니다.

따라서 쓰기 작업에는 읽기 작업보다 두 배 많은 SDS 대역폭이 필요합니다. 그러나 쓰기 작업에는 읽기 작업에 필요한 하나의 SDS가 아닌 두 개의 SDS가 사용됩니다. 따라서 읽기 및 쓰기 대역폭 요구 사항의 비율은 1:1.5입니다.

달리 말해 기본 스토리지의 처리량으로 비교하면 SDS당 쓰기 작업은 읽기 작업보다 1.5배 많은 네트워크 처리량을 필요로 합니다.

일반적인 상황에서는 앞서 설명한 스토리지 대역폭 계산으로 충분합니다. **그러나 환경에서 일부 SDS가 쓰기 작업이 많은 워크로드를 호스팅할 것으로 예상되는 경우 네트워크 용량을 추가하는 것이 좋습니다.**

7.2.3 볼륨이 다른 시스템으로 복제되는 환경

버전 3.5에서는 기본 비동기식 복제를 제공하므로 생성되는 대역폭을 고려할 때에는 첫째로 클러스터 내에서 생성되는 대역폭과 둘째로 복제본 피어 시스템 간에 생성되는 대역폭을 고려해야 합니다.

7.2.3.1 복제 시스템 내 대역폭

앞서 설명했듯이 볼륨이 복제되는 경우 I/O가 SDC에서 SDR로 전송되고, 이후에는 SDR에서 원본 시스템의 SDS로 후속 I/O가 전송됩니다. SDR은 먼저 볼륨 I/O를 연결된 SDS로 전달하여 처리(예: 압축)하고 디스크에 커밋합니다. 연결된 SDS는 SDR과 동일한 노드에 위치하지 않을 가능성이 있으므로 대역폭 계산 시 이를 고려해야 합니다. 두 번째 단계에서 SDR은 수신되는 쓰기를 저널링 볼륨에 적용합니다. 저널 볼륨은 PowerFlex 시스템 내의 다른 볼륨과 동일하기 때문에 SDR은 저널 볼륨이 존재하는 스토리지 풀을 지원하는 다양한 SDS에 I/O를 전송합니다. *이 단계에서는 SDR이 저널 볼륨을 지원하는 관련 기본 SDS에 먼저 작성하고 기본 SDS가 보조 SDS로 복제본을 전송하므로 두 개의 I/O가 더 추가됩니다.* 마지막으로, SDR은 원격 사이트에 전송하기 전에 저널 볼륨에서 추가 읽기를 수행합니다.

따라서 복제된 볼륨에 대한 쓰기 작업에는 복제되지 않은 볼륨에 대한 쓰기 작업보다 소스 클러스터 내에 대역폭이 3배 더 필요합니다. **복제된 볼륨에서 실행될 워크로드의 쓰기 프로파일을 신중하게 고려하십시오. 추가 쓰기 오버헤드를 수용하려면 추가 네트워크 용량이 필요합니다.** 따라서 시스템을 복제할 때는 백엔드 스토리지 트래픽을 수용할 수 있도록 4개의 25GbE 또는 2개의 100GbE 네트워크를 사용하는 것이 좋습니다.

7.2.3.2 복제본 피어 시스템 간 대역폭

복제본 피어 시스템 간의 네트워크 요구 사항을 고려하여 **소스 시스템과 타겟 시스템 간의 레이턴시가 200ms 이하여야 한다는 점이 거듭 강조됩니다.**

저널 데이터는 첫째, 복제 쌍 초기화 단계에서, 둘째, 복제 안정 상태 단계에서 소스 SDR과 타겟 SDR 간에 전송됩니다. LAN이든 WAN이든 소스 SDR과 타겟 SDR 간의 적절한 대역폭을 확보하도록 각별히 주의를 기울여야 합니다. WAN으로 연결하면 사용 가능한 대역폭을 초과할 가능성이 가장 큽니다. 쓰기 폴딩으로 타겟 저널에 전송될 데이터의 양을 줄일 수 있지만 항상 그렇게 된다고 쉽게 예측할 수 있는 사항이 아닙니다. *사용 가능한 대역폭이 초과되면 저널 간격이 백업을 하게 되면서 저널 볼륨 크기와 RPO가 모두 증가합니다.*

복제되는 모든 볼륨의 지속적인 쓰기 대역폭은 사용 가능한 총 WAN 대역폭의 80%를 넘지 않도록 하는 것이 가장 좋습니다. 피어 시스템 간에 볼륨을 상호 복제하는 경우 피어 SDR←→SDR 대역폭은 두 방향의 요구 사항을 동시에 고려해야 합니다. 특정 워크로드에 필요한 WAN 대역폭을 계산하는 데 도움이 필요하면 최신 PowerFlex Sizer를 참고하고 사용하십시오.

참고: Sizer 툴은 Dell Technologies 직원 및 파트너에게 제공되는 내부용 툴입니다. WAN 대역폭 사이징 지원이 필요한 외부 사용자는 기술 영업 전문가와 상의해야 합니다.

7.2.3.3 복제 상태에 대한 네트워킹 영향

이 백서에서는 PowerFlex 네트워킹 정보 모범 사례에 중점을 두고 있지만, 스토리지 계층 자체의 일반적인 운영, 상태 및 성능은 구축된 네트워크의 품질과 용량에 따라 달라집니다. 이는 비동기식 복제와 저널 볼륨의 사이징과 특별히 관련이 있습니다.

쓰기 최대치가 권장량인 "0.8 * WAN 대역폭"을 초과할 수 있지만 짧아야 합니다. 저널 크기는 이러한 쓰기 최대치를 흡수할 수 있을 만큼 커야 합니다.

이 점이 중요합니다. 저널 볼륨 용량은 피어 시스템 간 링크 중단을 감당할 수 있도록 크기가 조정되어야 합니다. 1시간의 운영 중단이 예상되더라도 사용자는 3시간을 기준으로 계획을 세워야 합니다. 하나는 운영 중단 시 애플리케이션 쓰기를 감안하여 충분한 저널 공간을 확보해야 합니다. **일반적으로 저널 용량은 최대 쓰기 대역폭 * 링크 다운타임으로 계산해야 합니다**. 가장 바쁜 시간 동안의 최대 애플리케이션 쓰기 대역폭을 알아야 합니다. 애플리케이션의 최대 쓰기 처리량이 초당 1GB라고 가정해보겠습니다. 3시간은 10,800초입니다. 따라서 필요한 저널 용량은

$$1\,GB/s * 10{,}800\text{초} = \sim 10.55\,TB$$

그러나 PowerFlex는 저널 용량을 풀 용량의 백분율로 설정합니다. 200TB 스토리지 풀 1개가 있다고 가정하는 경우:

$$100 * 10.55\,TB \,/\, 200\,TB = 5.27\%$$

안전 마진으로 이를 6%로 올림합니다.

참고: 저널 간격으로 전송된 볼륨 데이터는 압축되지 않습니다. PowerFlex에서 압축은 저장된 데이터용입니다. 잘 세분화된 스토리지 풀에서는 데이터가 SDC(복제되지 않은 볼륨의 경우) 또는 SDR(복제된 볼륨의 경우)에서 수신된 후 SDS 서비스에서 데이터 압축이 수행됩니다. SDR은 복제본 쌍의 양쪽에 있는 데이터 레이아웃을 인식하지 않으며 이에 구애받지 않습니다. 대상 또는 타겟 볼륨이 압축되어 구성된 경우 저널 간격이 적용될 때 타겟 시스템 SDS에서 압축이 수행됩니다.

7.2.4 하이퍼 컨버지드 환경

PowerFlex가 하이퍼 컨버지드로 구축된 경우 각 물리적 노드는 SDS, 하이퍼바이저의 SDC 및 하나 이상의 VM을 실행합니다. 이러한 의미에서 하이퍼 컨버지드 PowerFlex 구축 시에는 하이퍼바이저가 포함되지 않아도 됩니다. 하이퍼 컨버지드로 구축하면 하드웨어 투자를 최적화할 수 있지만, 네트워크 사이징도 필요로 합니다.

앞서 설명한 스토리지 대역폭 계산이 하이퍼 컨버지드 환경에 적용되지만, 모든 가상 머신, 하이퍼바이저 또는 OS 트래픽과 SDC에서 발생하는 트래픽에 대한 프런트엔드 대역폭도 고려해야 합니다. 가상 머신에 대한 사이징은 이 기술 보고서의 범위를 벗어나더라도 중요합니다.

하이퍼 컨버지드 환경에서는 다른 네트워크 트래픽과 논리적으로 별도의 스토리지를 제공하는 것도 중요합니다.

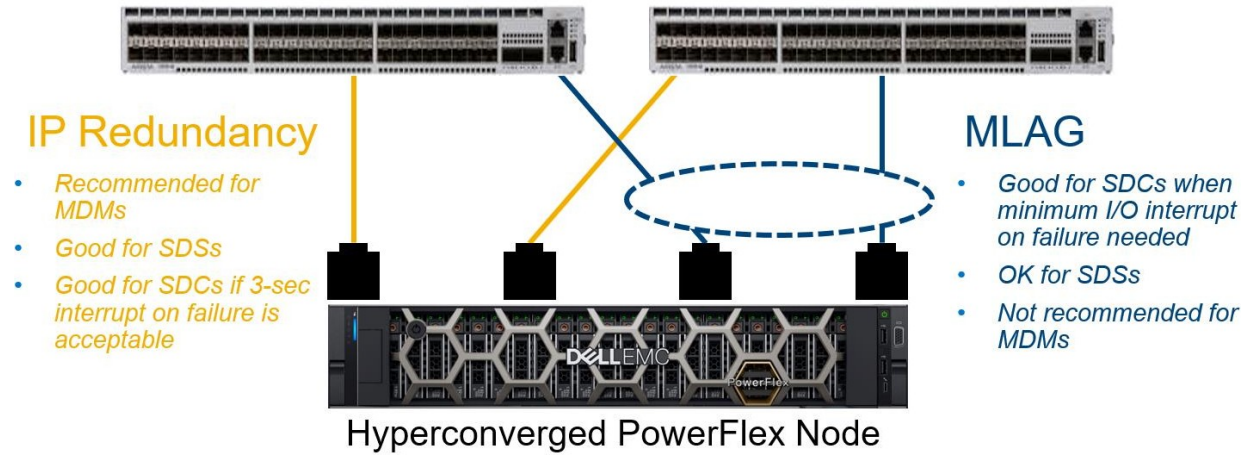

그림 8 4개의 25기가비트 네트워크 연결을 사용하는 하이퍼 컨버지드 VMware 환경의 예. 이 호스트의 PowerFlex 트래픽은 포트 EthO 및 Eth1을 활용합니다. 이중화는 MLAG가 아닌 기본 PowerFlex IP 경로 다중화와 함께 제공됩니다. 포트 Eth2 및 Eth3은 MLAG와 VLAN 태그 지정을 모두 사용하고 하이퍼바이저 및 다른 게스트에 대한 액세스 네트워크 액세스를 제공합니다. PowerFlex는 VLAN 태그 지정 및 링크 집선을 지원하므로 다른 구성이 가능합니다.

8 네트워크 하드웨어

8.1 전용 NIC

PowerFlex 엔지니어링 부서에서는 가급적 PowerFlex 트래픽 전용 네트워크 어댑터를 사용하는 것을 권장합니다. 전용 네트워크 어댑터는 전용 대역폭을 제공하고 문제 해결을 간소화합니다. 공유 네트워크 어댑터가 지원되며, 하이퍼 컨버지드 환경에서는 공유 네트워크 어댑터가 필수일 수 있습니다.

8.2 공유 NIC

최적의 옵션은 아니지만 PowerFlex 소프트웨어에서 공유 NIC를 사용할 수도 있습니다. PowerFlex 트래픽이 다른 비 PowerFlex 트래픽과 물리적 네트워크를 공유하는 경우, PowerFlex 또는 비 PowerFlex 트래픽으로 인해 네트워크 정체 또는 부족 문제가 발생하지 않도록 QoS를 구현해야 합니다.

8.3 2개의 NIC와 4개의 NIC 및 기타 구성 비교

PowerFlex를 사용하면 네트워크 인터페이스를 추가하여 네트워크 리소스를 확장할 수 있습니다. **필수는 아니지만 스토리지 네트워크의 프런트엔드 및 백엔드 트래픽을 격리시키는 것이 좋은 경우도 있습니다.** 이러한 방법은 스토리지 및 가상화 또는 컴퓨팅 팀이 각각 자체 네트워크를 관리하는 2계층 구축에 유용할 수 있습니다. 일반적으로 사용자는 프런트엔드 및 백엔드 네트워크 트래픽을 분할하여 스토리지 및 애플리케이션 관련 네트워크 트래픽의 성능을 보장할 수 있습니다. 모든 경우에 Dell Technologies에서는 이중화, 용량 및 속도를 위해 다수의 인터페이스를 권장합니다.

PCI NIC 이중화도 고려해야 합니다. **서버마다 2개의 듀얼 포트 PCI NIC를 사용하는 것이 1개의 쿼드 포트 PCI NIC를 사용하는 것보다 더 좋습니다. 단일 NIC의 2개의 듀얼 포트 PCI NIC를 구성하면 단일 NIC의 장애를 감당할 수 있기 때문입니다.**

8.4 스위치 이중화

대부분의 리프-스파인 구성에서, 스파인 스위치와 ToR(Top-of-Rack) 리프 스위치는 이중화됩니다. 이러한 방식은 ToR 스위치에 장애가 발생하는 경우 네트워크의 랙 내부의 구성 요소에 대한 지속적인 액세스를 제공합니다. 각 랙에 단일 ToR 스위치가 있으면 ToR 스위치 장애 발생 시 랙 내부의 SDS 구성 요소에 액세스할 수 없습니다. **따라서 단일 ToR 스위치 구성은 권장되지 않습니다.** 랙당 단일 ToR 스위치가 사용되는 경우 사용자는 스위치 장애 발생 시 데이터 가용성을 보장하기 위해 랙 수준에서 장애 세트를 정의해야 합니다.

9 IP 고려 사항

9.1 IPv4 및 IPv6

버전 2.6부터, PowerFlex는 3.0 이상의 모든 버전에서 2계층 구축 옵션과 하이퍼 컨버지드 구축 옵션 모두에 IPv6 지원을 제공합니다. 이전 버전의 PowerFlex는 IPv4(Internet Protocol version 4) 주소 지정만 지원합니다. 이 백서의 예에서는 IPv4에 중점을 둡니다.

9.2 IP 수준 이중화

MDM, SDS, SDR 및 SDC에 여러 개의 IP 주소를 지정할 수 있으므로 이들은 둘 이상의 네트워크에 상주할 수 있습니다. 이를 통해 로드 밸런싱과 이중화가 가능합니다.

PowerFlex는 기본적으로 소프트웨어 구성 요소가 여러 링크에서 트래픽을 보내도록 구성된 경우 물리적 네트워크 링크 전반에 이중화와 로드 밸런싱을 제공합니다. 이 구성에서 MDM, SDR 또는 SDS에 사용 가능한 각 물리적 네트워크 포트에는 각각 다른 서브넷에 고유한 IP 주소가 할당됩니다.

여러 서브넷을 사용하면 네트워크 수준에서 이중화가 제공됩니다. 또한 여러 서브넷을 사용하면 한 구성 요소에서 다른 구성 요소로 트래픽이 전송되는 경우 대상 IP 주소에 따라 소스 구성 요소의 라우팅 테이블에 있는 다른 항목이 선택됩니다. 이렇게 하면 소스가 단일 대상의 여러 IP 주소(각각 물리적 네트워크 포트에 대응)에 연결될 때 소스의 단일 물리적 네트워크 포트에서 발생하는 병목 현상을 방지할 수 있습니다.

달리 말하면 소스 및 대상의 여러 물리적 포트가 동일한 서브넷에 있는 경우 소스 포트에서 병목 현상이 발생할 수 있습니다. 예를 들어, 두 개의 SDS가 단일 서브넷을 공유하는 경우 각 SDS에는 물리적 포트가 두 개 있고, 각 물리적 포트의 서브넷에 고유 IP 주소가 할당되며, IP 스택으로 인해 소스 SDS는 항상 동일한 물리적 소스 포트를 선택합니다. **각 포트는 호스트의 라우팅 테이블에 있는 서로 다른 서브넷에 대응하므로 서브넷의 포트를 분할하면 로드 밸런싱이 가능합니다.**

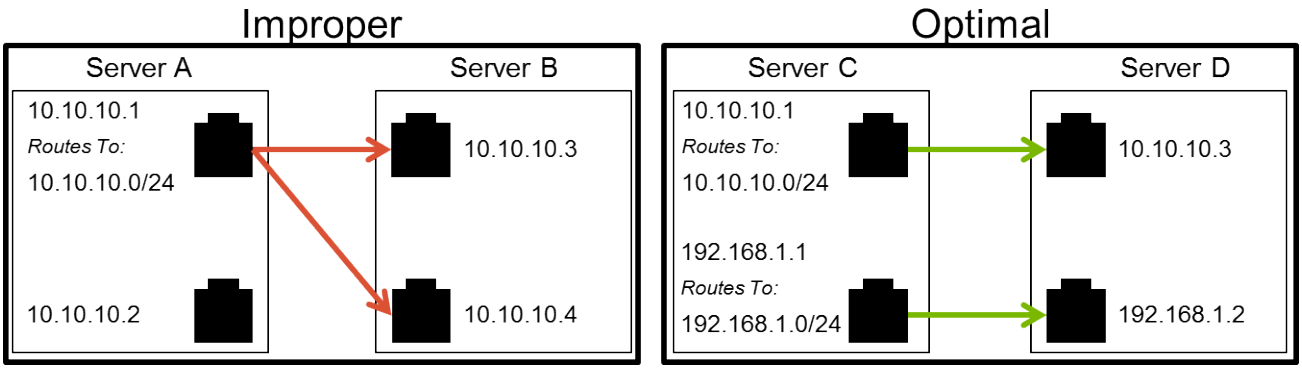

그림 9 운영 체제 IP 구성을 비교한 결과. 왼쪽의 부적절 한 IP 구성은 모든 트래픽에 동일한 서브넷 10.10.10.0/24를 사용합니다. 서버 A가 서버 B에 대한 연결을 시작하면 송신 연결에 항상 10.10.10.0/24로의 경로를 제공하는 네트워크 링크가 선택됩니다. 서버 A의 두 번째 네트워크 포트는 송신 연결에 사용되지 않습니다. 오른쪽의 올바른 IP 구성은 2개의 서브넷, 10.10.10.0/24 및 192.168.1.0/24를 사용하므로 서버 C의 두 포트를 모두 송신 연결에 사용할 수 있습니다. 참고: 이 예의 서브넷(10.10.10.0/24 및 192.168.1.0/24)은 임의로 선택한 것이며, 클래스 "A"와 클래스 "C"를 혼합하여 사용한 것은 시각적인 구별을 위해서입니다.

각 MDM 또는 SDS가 여러 IP 주소에 액세스할 수 있는 경우, PowerFlex는 트래픽 패턴을 인식하여 로드 밸런싱을 보다 효과적으로 처리합니다. 이를 통해 성능이 약간 향상될 수 있습니다.
또한 링크 집선은 링크 수준 페일오버에 대한 일련의 자체 타이머를 갖추고 있습니다. 따라서 링크 중단 시 기본적인 PowerFlex IP 수준 이중화로 문제를 쉽게 해결할 수 있습니다.

또한 IP 수준 이중화는 IP 주소 충돌을 방지합니다. 원치 않는 IP 변경이나 충돌을 방지하려면 **PowerFlex MDM 또는 SDC가 상주하는 네트워크에 DHCP를 구축해서는 안 됩니다.**

격리에 사용하는 경우 MDM 간 통신에 사용되는 링크에는 MLAG보다 IP 수준 이중화가 훨씬 좋습니다. IP 수준 이중화가 이중화 링크 집선 그룹에 기반한 VLAN에 계층화된 경우 두 기술을 모두 사용하는 것이 좋습니다. 이에 대한 예는 Dell EMC PowerFlex Appliance Network Planning Guide를 참조하십시오.

10 이더넷 고려 사항

10.1 점보 프레임

PowerFlex는 점보 프레임을 지원하며, 스토리지 트래픽에는 점보 프레임을 사용하는 것이 좋습니다. 하지만 네트워크 인프라스트럭처에 따라 점보 프레임을 사용하기 어려울 수 있습니다. 다양한 네트워크 구성 요소를 통해 점보 프레임을 일관성 없이 구현하면 해결하기 어려운 성능 문제가 발생할 수 있습니다. 점보 프레임을 사용하는 경우에는 호스트 및 스위치, 스토리지 VM(HCI가 구축된 경우) 등, PowerFlex 인프라스트럭처에서 사용하는 모든 네트워크 구성 요소에서 이를 사용해야 합니다.

점보 프레임을 사용하면 단일 이더넷 프레임에서 더 많은 데이터를 전달할 수 있습니다. 이렇게 하면 이더넷 프레임의 총 수와 각 노드에서 처리해야 하는 인터럽트 수가 줄어듭니다. PowerFlex 인프라스트럭처의 모든 구성 요소에서 점보 프레임을 사용하면 워크로드에 따라 약 10%의 성능상 이점이 있습니다.

참고: PowerFlex Manager를 사용하여 어플라이언스 또는 랙 시스템에 PowerFlex 클러스터를 구축하면 노드 및 스위치 구성 요소의 점보 프레임 구성이 모든 클러스터 구성 요소에 대해 완벽하게 조정 및 관리됩니다.

네트워크 구성 요소를 주의 깊게 검토하여 모든 지점에서 점보 프레임을 일관되게 구성하십시오. 확실치 않은 경우 처음에는 점보 프레임을 사용하지 않는 것이 좋습니다. 안정적으로 작동하는 상태에서 인프라스트럭처에 사용할 수 있는지 확인한 후에만 점보 프레임을 사용하십시오. 점보 프레임이 각 경로를 따라 모든 노드에 구성되도록 Linux `tracepath` 명령과 같은 유틸리티를 사용하여 경로를 따라 MTU 크기를 검색할 수 있습니다. Ping이 점보 프레임 문제를 진단하는 데 유용할 수 있습니다. Linux에서 `ping -M do -s 8972 <ip address/hostname>` 형태의 명령을 사용합니다. 여기에는 9,000MTU 크기에서 캡슐화되지 않은 패킷 헤더에 해당하는 28바이트가 빠져 있습니다.

점보 프레임 구현에 대한 자세한 내용은 PowerFlex Configure and Customize 가이드를 참조하십시오.

10.2 VLAN 태그 지정

PowerFlex는 서버 및 액세스 또는 리프 스위치 간 연결에 대한 기본 VLAN 및 VLAN 태그 지정에 구애받지 않습니다. 운영 체제 또는 스위치에서 구성되더라도 PowerFlex 소프트웨어에 영향이 없습니다. PowerFlex 엔지니어링 부서에서 측정했을 때 VLAN은 성능 수준에 영향을 미치지 않았습니다.

PowerFlex 어플라이언스 구축의 경우에는 동일한 VLAN 표준 세트를 구성해야 합니다. 아래 섹션 19를 참조하십시오.

11 링크 집선 그룹

LAG(Link Aggregation Group) 및 MLAG(Multi-Chassis Link Aggregation Group)는 엔드포인트 간 포트를 결합합니다. 엔드포인트는 스위치와 LAG 또는 2개의 스위치가 있는 호스트, 그리고 MLAG가 있는 호스트가 될 수 있습니다. 링크 집선 용어 및 구현 방식은 스위치 공급업체에 따라 다릅니다. Cisco Nexus 스위치의 MLAG 기능은 vPC(Virtual Port Channel)라고 합니다.

LAG는 설정, 연결 해제 및 오류 처리에 LACP(Link Aggregation Control Protocol)를 사용합니다. LACP는 표준이지만 다양한 독자적 변형이 있습니다.

스위치 공급업체나 PowerFlex를 호스팅하는 운영 체제와 관계없이 **링크 집선 그룹을 사용하는 경우에는 LACP가 권장됩니다. 정적 링크 집선은 사용할 수 없습니다.**

물리적 포트마다 고유한 IP 주소가 지정되는 IP 수준 이중화를 대신해서 링크 집선을 사용할 수 있습니다. 링크 집선은 일부 팀의 경우 구성하기 더 간단할 수 있으며 IP 주소 소진이 문제가 되는 상황에서 유용합니다. PowerFlex를 실행하는 노드, 그리고 연결되는 네트워크 장비 모두에서 링크 집선을 구성해야 합니다.

PowerFlex는 IP 수준 이중화 또는 링크 집선 선택 여부와 관계없이 뛰어난 회복탄력성과 성능을 발휘합니다. MLAG를 사용하는 경우 SDS의 성능은 IP 수준 이중화의 성능에 근접합니다.

- **SDS에 대한 MLAG 또는 IP 수준 이중화 선택 여부는 운영상의 결정으로 간주되어야 합니다.**
- **MDM 간 트래픽의 경우 MDM에서 하나의 IP 주소를 지속적으로 사용하여 여러 IP 주소 간에 통신하도록 설계된 MDM 간의 짧은 시간 초과로 인한 페일오버를 방지하는 데 도움이 되므로 MLAG보다 IP 수준 이중화 또는 LAG를 사용하는 것이 좋습니다.**
- **3.5의 네트워크 장애 회복탄력성이 개선되었기 때문에 SDC 구성 요소에서 사용 중인 링크의 경우 일반적으로 IP 수준 이중화가 MLAG보다 선호됩니다.**

11.1 LACP

LACP는 정기적으로 네트워크 링크 집선 그룹의 각 물리적 네트워크 링크에 메시지를 전송합니다. 이 메시지는 각 물리적 링크가 여전히 활성 상태인지 확인하는 논리의 일환입니다. 이러한 메시지의 빈도는 네트워크 관리자가 LACP 타이머를 사용하여 제어할 수 있습니다.

LACP 타이머는 일반적으로 빠른 속도(초당 메시지 1개) 또는 정상 속도(30초마다 메시지 1개)로 링크 장애를 탐지하도록 구성할 수 있습니다. LACP 타이머가 빠른 속도로 작동하도록 구성하면 개선 조치가 신속하게 이루어집니다. 또한, 최신 네트워크 기술을 사용하면 초당 메시지 전송에 대한 상대적인 오버헤드가 작습니다.

PowerFlex SDS와 스위치 사이에 링크 집선을 사용할 때에는 LACP 타이머가 빠른 속도로 작동하도록 구성해야 합니다.

LACP 연결을 설정하려면 LACP 피어 중 하나 또는 둘 다 활성 모드를 사용하도록 구성해야 합니다. **따라서 PowerFlex 노드에 연결된 스위치는 링크에서 활성 모드를 사용하도록 구성하는 것이 좋습니다.**

11.2 로드 밸런싱

링크 집선 그룹에서 여러 네트워크 링크가 활성인 경우 엔드포인트는 링크 사이에 트래픽을 분산하는 방법을 선택해야 합니다. 네트워크 관리자는 엔드포인트에 로드 밸런싱 방식을 구성하여 이 동작을 제어합니다. 로드 밸런싱 방식은 일반적으로 소스 또는 대상 IP 주소, MAC 주소 또는 TCP/UDP 포트의 조합에 따라 사용할 네트워크 링크를 선택합니다.

이 로드 밸런싱 방식은 "해시 모드"라고 부릅니다. 해시 모드 로드 밸런싱은 링크가 활성 상태로 유지되는 경우 동일한 물리적 링크의 특정 소스 및 대상 주소/전송 포트 쌍에서 오가는 트래픽을 유지하는 것을 목표로 합니다.

해시 모드 로드 밸런싱의 권장 구성은 사용 중인 운영 체제에 따라 달라집니다.

SDS를 실행하는 노드에 스위치에 대한 집선 링크가 있고, 이 노드가 VMware ESX®를 실행하는 경우 "소스 및 대상 IP 주소" 또는 "소스 및 대상 IP 주소 및 TCP/UDP 포트"를 사용하도록 해시 모드를 구성해야 합니다.

SDS를 실행하는 노드에 스위치에 대한 집선 링크가 있고 이 노드가 Linux를 실행하는 경우 "xmit_hash_policy=layer2+3" 또는 "xmit_hash_policy=layer3+4" 본딩 옵션을 사용하도록 Linux의 해시 모드를 구성해야 합니다. "xmit_hash_policy=layer2+3" 본딩 옵션은 로드 밸런싱에 소스 및 대상 MAC 주소와 IP 주소를 사용합니다. "xmit_hash_policy=layer3+4" 본딩 옵션은 로드 밸런싱에 소스 및 대상 IP 주소와 TCP/UDP 포트를 사용합니다.

Linux에서는 "miimon=100" 본딩 옵션도 사용해야 합니다. 이 옵션은 Linux가 100밀리초마다 각 물리적 링크의 상태를 확인하도록 지시합니다.

각 본딩 옵션의 이름은 Linux 배포 환경에 따라 다를 수 있지만 권장 사항은 동일합니다.

11.3 MLAG(Multiple Chassis Link Aggregation Group)

MLAG는 LAG(Link Aggregation Group)와 마찬가지로 네트워크 링크 이중화를 제공합니다. LAG와는 달리, MLAG는 단일 엔드포인트(예: PowerFlex를 실행하는 노드)를 여러 스위치에 연결할 수 있습니다. 스위치 공급업체는 MLAG를 칭할 때 다른 이름을 사용하며, MLAG 구현 방식은 대개 독자적입니다.

PowerFlex에서 MLAG를 사용할 수 있지만 일반적으로 MDM 간 트래픽에는 MLAG를 권장하지 않습니다. 단 다음 섹션의 참고 사항을 확인하십시오. "로드 밸런싱" 섹션에 설명된 옵션에는 MLAG도 적용됩니다.

12 MDM 네트워크

MDM은 호스트(SDC)와 분산 스토리지(SDS) 간의 데이터 경로에 상주하지 않지만 클러스터 상태를 지속적으로 확인하기 위해 서로 간의 관계를 유지해야 합니다. 따라서 MDM 간 트래픽은 MLAG의 물리적 네트워크 링크 손실과 같이 레이턴시에 영향을 주는 네트워크 이벤트에 민감합니다.

MDM은 이중화됩니다. 따라서 PowerFlex는 레이턴시 증가뿐만 아니라 MDM 손실도 감당할 수 있습니다. MDM을 호스팅하는 노드에 MLAG를 사용해도 됩니다. **그러나 MDM 간 트래픽을 전달하는 네트워크에서 MLAG를 사용해야 하는 경우 Dell EMC PowerFlex 담당자와 협력하여 MLAG와 기본 IP 레벨 이중화를 함께 사용하여 두 가지 네트워크 이중화를 제공하는 강력한 설계를 채택하십시오.**

대부분의 경우, MDM에 MLAG보다는 두 개 이상의 네트워크 세그먼트에 기반한 IP 수준 이중화를 사용하는 것이 좋습니다. MDM은 하나 이상의 전용 MDM 클러스터 네트워크를 공유할 수 있습니다.

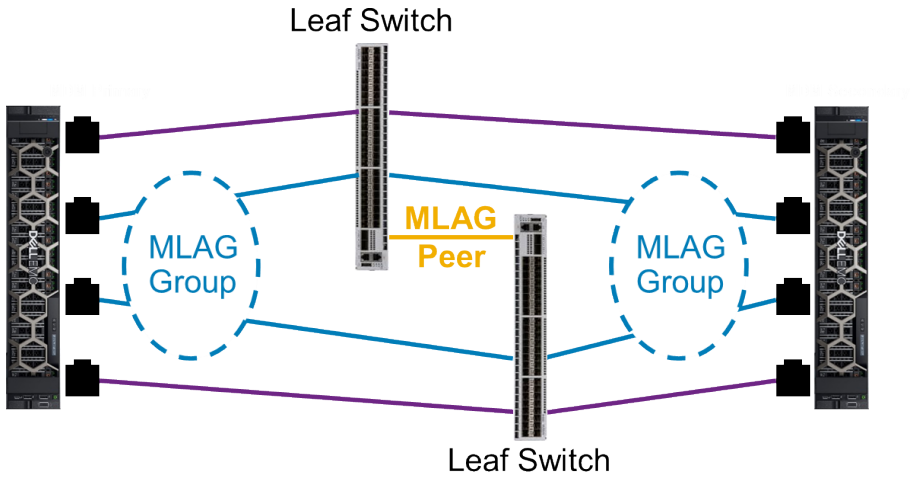

그림 10 2개의 리프 스위치에 연결된 노드 2개. MDM 트래픽은 MLAG 그룹에 있지 않기 때문에 보라색 링크를 통과해야 합니다.

13 네트워크 서비스

13.1 DNS

MDM 클러스터는 시스템 구성 요소와 해당 IP 주소의 데이터베이스를 유지 관리합니다. PowerFlex 구축에 영향을 미치는 DNS 중단이 발생할 가능성을 없애기 위해 MDM 클러스터는 호스트 이름 또는 FQDN(Fully Qualified Domain Name)을 통해 시스템 구성 요소를 추적하지 않습니다. MDM 클러스터와 함께 시스템 구성 요소를 등록할 때 호스트 이름 또는 FQDN을 사용하는 경우, 이는 IP 주소로 확인되고 구성 요소가 해당 IP 주소로 등록됩니다.

예외 사항은 VASA 공급자가 배포되고 vVols가 구현되는 경우입니다. PowerFlex 환경에서 vVols를 사용하려면 단일 모드 또는 3노드 클러스터 중 하나로 PowerFlex VASA 공급자를 배포해야 합니다. vVols 기술을 vSphere 환경에 구현하려면 vCenter Server, vVol 데이터스토어를 사용할 ESXi 호스트 및 VASA 공급자 자신의 FQDN이 모두 필요합니다.
이러한 모든 구성 요소에 대해 유효한 DNS 확인이 이루어져야 합니다. 따라서 DNS 서비스는 가용성이 높아야 vVol 연결 및 기능 손실을 방지할 수 있습니다.

요약하자면 **vVols가 구현되지 않는 한, PowerFlex 구축 환경에서는 일반적으로 호스트 이름 및 FQDN 변경이 구성 요소 간 트래픽에 영향을 미치지 않습니다.**

14 WAN을 통한 복제 네트워크

PowerFlex 기본 비동기식 복제를 사용할 때 고려해야 할 추가 고려 사항이 있습니다. 섹션 2.4와 3.9에서는 SDR(Storage Data Replicator)과 해당 트래픽을 다루었습니다. 섹션 7.2.3에서는 추가 대역폭 요구 사항을 다루었습니다. 이 섹션에서는 WAN(Wide Area Network)을 통한 복제를 실행하는 작업과 관련하여 주소 지정과 라우팅을 고려합니다. 구현 세부 사항은 사용하는 하드웨어 및 WAN 토폴로지에 따라 달라지므로 일반적인 권장 사항을 다룹니다.

14.1 추가 IP 주소

보호 도메인 내에서 SDR은 SDS와 동일한 호스트에 설치되지만 SDR이 저널 볼륨에 작성하는 트래픽은 호스트에 공동 위치하는 SDS뿐만 아니라 저널을 호스팅하는 모든 SDS로 전송됩니다. 백엔드 스토리지 네트워크에서 각 SDR은 SDS와 동일한 노드 IP에서 수신 대기하므로 보호 도메인의 모든 SDS에 연결할 수 있어야 합니다.

그러나 SDR에는 원격 SDR과 통신할 수 있는 별개의 추가 IP 주소가 필요합니다. 대부분의 경우 적절하게 구성된 게이트웨이로 라우팅할 수 있는 IP 주소여야 합니다. 이중화를 위해 각 SDR에 2개의 IP 주소가 필요합니다.

14.2 방화벽 고려 사항

SDR은 서로 통신하며, TCP 포트 1088을 통해 복제된 데이터를 서로 간에 전송합니다. 이 포트는 소스 시스템 측 방화벽에서 송신용으로 열려 있어야 하며, 타겟 시스템 측의 수신용으로 열려 있어야 합니다. 두 시스템 간 양방향으로 복제가 이루어지는 경우 양 측의 송신 및 수신 모두를 위해 포트 1088이 방화벽에서 열려 있어야 합니다.

14.3 정적 라우팅

PowerFlex 비동기식 복제는 일반적으로 동일한 주소 세그먼트를 공유하지 않는 물리적 원격 클러스터 간에 WAN을 통해 발생합니다. 기본 경로 자체가 원격 SDR IP로 패킷을 적절히 전송하는 데 적합하지 않은 경우, 고정 경로는 다음 홉 주소나 송신 인터페이스 중 하나 또는 원격 서브넷에 연결하기 위해 두 가지 모두를 나타내도록 구성해야 합니다.

예: `X.X.X.X/X via X.X.X.X dev interface`

각 측면에 몇 개의 노드가 있는 소규모 시스템을 생각해 봅니다. 각 노드에는 4개의 네트워크 어댑터가 있으며, 2개는 PowerFlex 클러스터의 내부 통신을 위해 IP로 구성되고, 2개는 사이트 간, 외부 통신을 위해 IP 주소로 구성됩니다.

이 예에서는 지정된 게이트웨이를 통해 다른 쪽의 WAN 서브넷에 액세스하도록 노드에 지시합니다. 소스 사이트 A에서 네트워크 인터페이스 `enp130s0f0` 및 `enp130s0f1`은 각각 `30.30.214.0/24`와 `32.32.214.0/24` 범위의 주소로 구성됩니다. 각각에 라우트 인터페이스 파일을 구성하여 지정된 게이트웨이 및 인터페이스를 통해 원격 네트워크용 패킷을 전송할 수 있습니다.

```
route-enp130s0f0 내용 →          31.31.0.0/16 via 30.30.214.252 dev enp130s0f0
route-enp130s0f1 내용 →          33.33.0.0/16 via 32.32.214.252 dev enp130s0f1
```

원격 네트워크 31.31.214.0/24로 향하는 패킷은 게이트웨이 IP 30.30.214.252에서 다음 홉 주소를 통해 전송됩니다. 33.33.214.0/24로 향하는 패킷에 대해서도 마찬가지입니다.

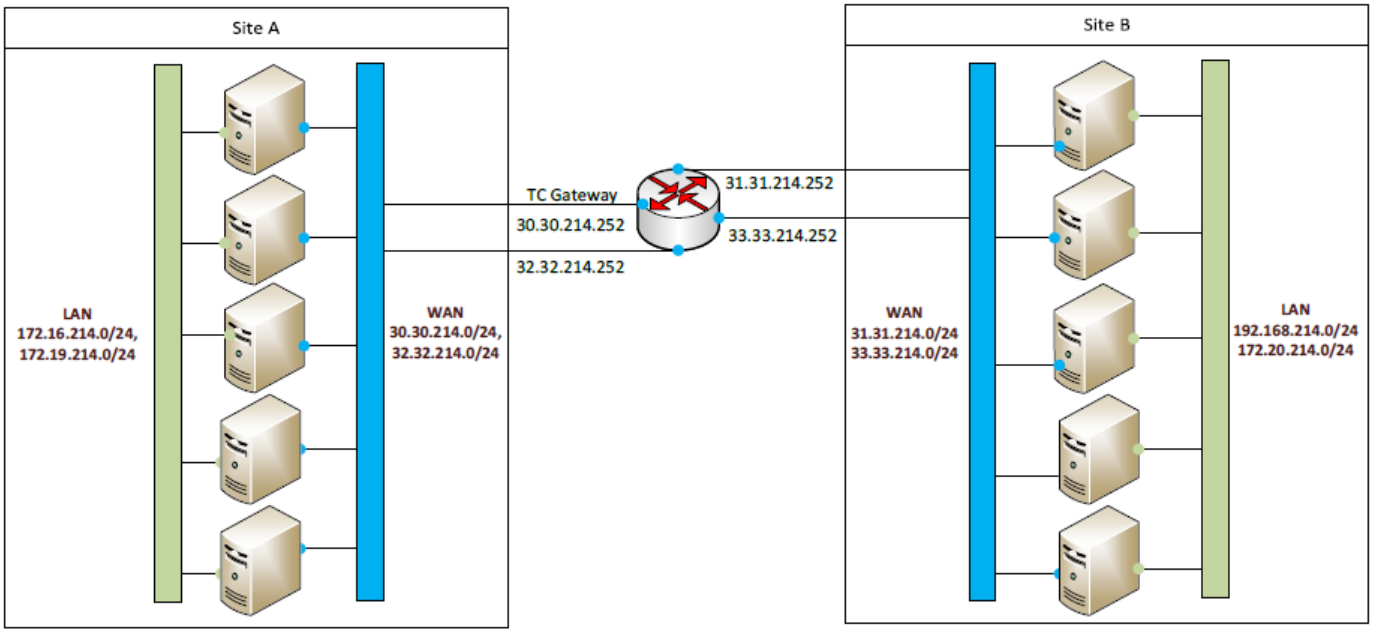

그림 11 PowerFlex 복제를 위한 WAN 토폴로지 예.

정적 라우팅 구성에 대한 세부 사항은 운영 체제/하이퍼바이저 및 전체 네트워크 아키텍처에 따라 다르지만 일반적인 원칙은 동일합니다.

14.4 MTU 및 점보 프레임

MTU는 WAN 링크 구성과 일치하도록 상호 SDR 네트워크 인터페이스에서 올바르게 설정되어야 합니다. 대부분의 경우 이는 1500입니다. 이는 성능 향상을 위해 모든 로컬 네트워크에서 점보 프레임을 사용하는 경우 특히 기억해야 할 중요한 요소입니다. MTU가 WAN 구성과 일치하지 않는 경우 IP 조각화로 인해 복제 성능이 저하됩니다. 하드웨어 구성에 따라 MTU가 일치하지 않는 경우 인터페이스에 도달할 때 패킷이 모두 손실될 수 있습니다. 따라서 모든 경우에 WAN의 MTU를 확인하고 테스트해야 합니다.

15 동적 라우팅 고려 사항

수백 개의 노드로 구성된 대규모 리프 스파인 환경에서는 네트워크 인프라스트럭처에 PowerFlex 트래픽의 동적 라우팅이 필요할 수 있습니다.

PowerFlex 트래픽의 라우팅하는 핵심 목적은 라우팅 프로토콜의 컨버전스 시간을 단축하는 것입니다. 구성 요소나 링크에 장애가 발생하면 라우터나 스위치가 오류를 탐지해야 하고, 라우팅 프로토콜은 변경 사항을 다른 라우터에 전파해야 하고, 각 라우터 또는 스위치는 각 대상 노드의 라우팅을 다시 계산해야 합니다. 네트워크가 올바르게 구성된 경우 이 프로세스는 300밀리초 미만으로 발생할 수 있으며, 이는 MDM 클러스터 안정성을 유지하기에 충분한 속도입니다.

심한 정체 또는 네트워크 장애가 발생하여 컨버전스 시간이 400밀리초를 초과하면 MDM 클러스터가 보조 MDM으로 페일오버할 수 있습니다. **300밀리초가 시스템 안정성을 최대로 유지하기 위한 목표치**이기는 하지만 MDM이 페일오버하면 시스템이 계속 작동하고, I/O가 계속 이루어집니다.
다른 시스템 구성 요소 통신 메커니즘의 시간 제한 값이 훨씬 더 높으므로, 시스템은 가장 까다로운 시간 제한 요구 사항(MDM의 요구 사항)을 기준으로 설계되어야 합니다.

가능한 가장 빠른 컨버전스 시간을 위해 표준 모범 사례가 적용됩니다. 이는 신속한 컨버전스를 막는 불충분한 라우터(약한 링크)를 배제하는 일을 포함하여 최종 목표를 달성하기 위해 마련된 모든 네트워크 공급업체 모범 사례를 준수한다는 의미입니다.

테스트된 모든 네트워크 공급업체의 기본 OSPF 또는 BGP 구성에서는 컨버전스 시간이 충분하지 않습니다. **네트워크 공급업체와 관계없이 라우팅 프로토콜을 구축할 때마다 컨버전스 시간을 최소화하기 위한 성능 미세 조정 작업이 수반되어야 합니다.** 이러한 미세 조정에는 BFD(Bidirectional Forwarding Detection)의 사용과 장애 관련 타이밍 메커니즘의 조정이 포함됩니다.

OSPF와 BGP는 모두 PowerFlex 테스트를 거쳤습니다. PowerFlex는 라우팅 프로토콜과 네트워킹 디바이스가 올바르게 구성된 경우 링크 및 디바이스 장애 시에도 오류 없이 작동하는 것으로 알려져 있습니다. 단, **BGP보다는 OSPF가 권장됩니다**. 둘 다 빠른 컨버전스에 맞게 최적으로 구성된 경우 BGP보다 OSPF의 컨버전스가 빠른 것으로 나타난 테스트 결과가 이 권장 사항을 뒷받침합니다.

15.1 BFD(Bidirectional Forwarding Detection)

라우팅 프로토콜 옵션(OSPF 또는 BGP)과 관계없이 BFD(Bidirectional Forwarding Detection) 사용은 필수입니다. BFD는 프로토콜 기본 Hello 타이머와 관련된 오버헤드를 줄여 링크 오류를 신속하게 탐지할 수 있습니다. BFD는 라우터 CPU 및 대역폭을 적게 사용하는 등의 여러 가지 이유로 기본 프로토콜 Hello 타이머보다 더 빠른 오류 탐지를 제공합니다. **따라서 프로토콜 Hello 타이머를 주력으로 사용하는 것보다 BFD를 사용하는 것이 좋습니다.**

PowerFlex는 BFD와 최적화된 OSPF 및 BGP 라우팅을 사용하여 구축되었을 때 네트워크 페일오버 시 안정적입니다. 1초 미만으로 오류를 탐지하려면 BFD를 사용해야 합니다.

네트워크 컨버전스를 위해서는 이벤트를 탐지한 후 다른 라우터로 전파하고 해당 라우터가 이를 처리해야 하며, RIB(Routing Information Base) 또는 FIB(Forwarding Information Base)를 업데이트해야 합니다. 컨버전스 대상 라우팅 프로토콜에서 이 모든 단계가 이루어져야 하며, 해당 단계는 모두 300밀리초 미만으로 완료되어야 합니다.

Cisco 9000 Series 스위치를 사용한 테스트에서는 **150밀리초의 BFD Hold Down 타이머**로 충분했습니다. 150밀리초 Hold Down 타이머의 구성은 50밀리초의 min_rx와 3의 승수를 사용한 50밀리초의 전송 간격으로

구성됩니다. PowerFlex 권장 사항은 최대 150밀리초의 Hold Down 타이머를 사용하는 것입니다. 스위치 공급업체가 150밀리초 미만의 BFD Hold Down 타이머를 지원하는 경우 달성 가능한 가장 짧은 Hold Down 타이머를 사용하는 것이 좋습니다. 가능한 경우 비동기식 모드로 BFD를 사용해야 합니다.

Cisco vPC(MLAG)를 사용하는 환경에서는 **FHRP(First Hop Redundancy Protocol)를 실행하는 모든 라우팅 인터페이스와 모든 호스트용 인터페이스에서도 BFD를 사용하도록 설정해야 합니다.**

```
feature bfd

hsrp bfd all-interfaces

interface Vlan<num>
no shutdown
no ip redirects
ip address 192.168.103.2/24
no ipv6 redirects
hsrp version 2
hsrp 103
authentication text Vce12345
preempt
priority 110
ip 192.168.103.1

router ospf 1
  bfd
  bfd all-interfaces strict-mode

interface eth <x/x> / vlan <num> / Po <num>
 bfd interval 50 min_rx 50 multiplier 3
```

그림 12 집선 - 액세스/스파인 리프 토폴로지를 사용하는 Cisco 스위치의 BFD 구성 예. BFD는 150밀리초의 Hold Down 타이머를 사용하여 구성됩니다(간격은 50마이크로초, 승수는 3). 인터페이스 port-channel51의 OSPF와 인터페이스 Vlan30의 HSRP는 모두 BFD 클라이언트로 구성됩니다.

```
bfd ipv4 interval 50 min_rx 50 multiplier 3


interface Vlan30
  bfd interval 50 min_rx 50 multiplier 3
  no bfd echo
  vrrp 1
    vrrp bfd 30.30.30.124

interface port-channel49
  no bfd echo
  bfd per-link

interface port-channel51
  no bfd echo
  bfd per-link
router ospf 100
  bfd
```

그림 13 집선 - 액세스 토폴로지의 Dell BCFD 구성의 예. BFD는 150밀리초의 Hold Down 타이머를 사용하여 구성됩니다(간격은 50마이크로초, 승수는 3). 인터페이스 port-channel51의 OSPF와 인터페이스 Vlan30의 VRRP는 모두 BFD 클라이언트로 구성됩니다.

이 구성에 대한 다음 내용을 참조하십시오.

- port-channel 인터페이스의 경우, 링크별로 BFD를 사용해야 합니다.
- BFD에서 IP 리디렉션을 사용해서는 안 되며 BFD가 작동하도록 재정의해야 합니다.
- FHRP는 액세스/집선 토폴로지에만 필요합니다.

15.2 물리적 링크 구성

링크 오류와 관련된 타이머는 튜닝 대상입니다. 링크 중단/인터페이스 중단 이벤트 탐지 및 처리는 네트워크 공급업체 및 제품군에 따라 다릅니다. **Cisco Nexus 스위치에서 "`carrier-delay`" 타이머는 각 인터페이스에서 100밀리초로 설정해야 하며 "`link debounce`" 타이머는 각 물리적 인터페이스에서 500밀리초로 설정해야 합니다.**

반송파 지연(`carrier-delay`)은 스위치의 타이머이며 SVI 인터페이스에 적용 가능합니다. 반송파 지연은 스위치에서 링크 오류가 탐지될 때 애플리케이션에 알리기 전에 대기해야 하는 시간을 나타냅니다. 반송파 지연은 불안정한 네트워크에서 플래핑 이벤트 알림을 방지하는 데 사용됩니다. 최신 리프 스파인 환경에서는 모든 링크를 P2P(Point to Point)로 구성하여 안정적인 네트워크를 제공해야 합니다. PowerFlex 트래픽을 전달하는 SVI 인터페이스에 대한 권장 값은 100밀리초입니다.

디바운스(`link Debounce`)는 펌웨어에서 링크 중단 알림을 지연하는 타이머이며 물리적 인터페이스에 적용할 수 있습니다. 디바운스는 반송파 지연과 비슷하지만 논리적 인터페이스가 아닌 물리적 인터페이스에 적용 가능하며 링크 중단 알림에만 사용됩니다. 대기 기간에는 트래픽이 중지됩니다. 0이 아닌 링크 디바운스 설정은 라우팅 프로토콜의 컨버전스에 영향을 줄 수 있습니다. 링크 디바운스 타이머의 권장 값은 PowerFlex 트래픽을 전달하는 물리적 인터페이스에 대해 500밀리초입니다.

```
interface vlan <num>
  carrier-delay msec 100

interface eth <x/x>
  link debounce time 500
```

15.3 ECMP

ECMP(Equal-Cost Multi-Path Routing)를 사용해야 합니다. ECMP는 리프 및 스파인 스위치 사이에 트래픽을 균등하게 분산하고 이중화된 리프-스파인 네트워크 링크를 사용하여 고가용성을 제공합니다. ECMP는 MLAG와 유사하지만 이더넷이 아닌 Layer 3(IP)을 통해 작동합니다.

ECMP는 Cisco Nexus 스위치에 OSPF를 사용하는 경우 기본적으로 설정되어 있습니다. Cisco Nexus 스위치에 BGP를 사용하는 경우에는 기본적으로 설정되어 있지 않으므로 수동으로 활성화해야 합니다. 사용되는 ECMP 해시 알고리듬은 Layer 3(IP) 또는 Layer 3 및 Layer 4(IP 및 TCP/UDP 포트)여야 합니다.

15.4 OSPF

OSPF는 올바르게 구성된 경우 신속한 컨버전스가 가능하므로 선호되는 라우팅 프로토콜입니다. OSPF를 사용하는 경우 리프 및 스파인 스위치가 모두 단일 OSPF 영역에 상주합니다. **안정적인 MDM 간 통신을 제공하려면 300밀리초 미만의 컨버전스 시간이 필요합니다.** 모든 리프 및 스파인 스위치에서 OSPF 인터페이스는 P2P(Point to Point)로 구성되어야 하며, 이때 OSPF 프로세스는 BFD의 클라이언트로 구성됩니다.

이렇게 하면 타이머가 올바르게 설정되며 기본값에는 차이가 없습니다. **또한 ToR-집선(액세스-집선) 토폴로지를 사용한 L3 전달 방식에서는 OSPF 인터페이스를 P2P(Point to Point)로 구성해야 합니다.**

15.5 BGP

OSPF는 더 빠른 컨버전스가 가능하기 때문에 선호되지만, BGP도 필요한 시간 내에 컨버전스가 가능하도록 구성할 수 있습니다.

BGP는 기본적으로 Cisco Nexus 스위치에서 ECMP를 사용하도록 구성되지 않으므로 수동으로 구성해야 합니다. IBGP와 EBGP는 모두 기본적으로 ECMP를 지원하지 않으므로 별도 구성해야 합니다. IBGP의 구성에는 스파인-리프 토폴로지에서 ECMP를 완벽하게 지원할 수 있도록 BGP 라우팅 리플렉터와 경로 추가 기능이 필요합니다.

각각의 리프 및 스파인 스위치가 서로 다른 ASN(Autonomous System Number)을 나타내는 방식으로 BGP를 구성할 수 있습니다. 이 구성에서는 리프마다 다른 스파인과 피어를 이루어야 합니다.

또한 리프 및 스파인 스위치는 스위치가 여러 BGP 경로에서 로드 밸런싱할 수 있도록 ECMP를 설정해야 합니다. Cisco에서는 여기에 "`maximum-path`" 매개변수를 스파인 스위치에 사용 가능한 경로 수로 설정하는 작업이 포함됩니다.

PowerFlex에 BGP를 사용하려면 각 리프 및 스파인 이웃에 BFD를 구성해야 합니다. BGP를 사용하는 경우 **SDS 및 MDM 네트워크는 리프 스위치에 의해 알려집니다.**

Leaf Configuration

```
router bgp 100
  router-id 1.1.1.2
  address-family ipv4 unicast
    maximum-paths ibgp 3
    address-family l2vpn evpn
    maximum-paths ibgp 3

  neighbor 11.11.11.11
    bfd
    remote-as 100
    update-source loopback0
    address-family ipv4 unicast
      send-community
      send-community extended
    address-family l2vpn evpn
      send-community
      send-community extended

  vrf VxFLEX_MGMTanagement_VRF
    address-family ipv4 unicast
      maximum-paths ibgp 3
      advertise l2vpn evpn
      redistribute direct route-map ALL
```

Spine Configuration

```
router bgp 100
  router-id 11.11.11.11
  address-family ipv4 unicast
   maximum-paths ibgp 3
  address-family l2vpn evpn
   maximum-paths ibgp 3

  neighbor 1.1.1.1
    bfd
    remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community
      send-community extended
      route-reflector-client
```

그림 14 Cisco Nexus 리프 스위치(왼쪽) 및 스파인 스위치(오른쪽)에 대한 BGP 구성 예. 이들은 동일한 자율 운영 시스템(100)에 상주합니다. "maximum-path" 매개변수는 ECMP에 사용할 경로 수에 맞게 조정됩니다. 이 예에서는 3이지만 아닌 경우도 있습니다. 각 리프 또는 스파인 이웃에 BFD를 사용합니다. 리프 스위치는 PowerFlex MDM 및 SDS 네트워크를 알리기 위해 구성되었습니다.

참고:

- 스파인-리프 토폴로지를 사용하는 PowerFlex 랙 시스템에서는 컨트롤 플레인의 통신과 EVPN의 도달력을 위해 BGP가 사용됩니다. OSPF는 데이터 플레인에 사용됩니다.
- maximum-paths는 여러 NVE 인터페이스의 VTEP 도달력을 결정합니다.
- IBGP는 라우팅 리플렉터로 스파인을 사용하도록 구성됩니다
- EBGP를 사용하지 않으므로 BGP에 as-path multipath-relax를 적용할 수 없습니다.

15.6 리프-스파인 대역폭 요구 사항

스토리지 미디어가 성능 병목 현상을 일으키지 않는다고 가정하고, 리프 스위치와 스파인 스위치 간에 필요한 대역폭 양을 계산하려면 각 리프 스위치에서 연결된 호스트까지 사용 가능한 대역폭 양을 확인하고 리프 스위치에 대해 로컬일 가능성이 높은 I/O를 차감한 다음 각 스파인 스위치 간의 원격 대역폭 요구량을 나누어야 합니다.

랙마다 리프 스위치 2개와 서버 20대가 있고, 각 서버에는 25기가비트 인터페이스가 2개 있으며, 각 서버는 랙의 리프 스위치 2개에 이중으로 연결된 경우를 가정해봅니다. 이 경우 각 리프 스위치의 다운스트림 대역폭은 다음과 같이 계산됩니다.

$$20\ servers * 25\ \frac{\text{기가비트}}{server} = 500\ \text{기가비트}$$

각 리프 스위치에 대한 다운스트림 대역폭 요구 사항은 500기가비트입니다. 그러나 일부 트래픽은 리프 스위치 쌍에서 로컬이 되므로 스파인 스위치를 통과할 필요가 없습니다.

랙의 리프 스위치에 로컬인 트래픽 양은 구성의 랙 수에 따라 결정됩니다. 2개의 랙이 있는 경우, 트래픽의 50%가 로컬일 가능성이 높습니다. 3개의 랙이 있는 경우, 트래픽의 33%가 로컬일 가능성이 높습니다. 4개의 랙이 있는 경우, 트래픽의 25%가 로컬일 가능성이 높습니다. 달리 말해, 원격일 가능성이 있는 I/O의 비율은 다음과 같습니다.

$$remote\_ratio = \frac{number\_of\_racks - 1}{number\_of\_racks}$$

이 예에서는 2개의 랙이 있으므로 대역폭의 50%가 원격일 가능성이 높습니다.

$$remote\_ratio = \frac{2\ total\_racks - 1\ rack}{2\ total\_racks} = 50\%$$

이 예에서는 2개의 랙이 있으므로 대역폭의 50%가 원격일 가능성이 높습니다. 원격일 것으로 예상되는 트래픽 양에 각 리프 스위치의 다운스트림 대역폭을 곱하면 각 리프 스위치의 총 원격 대역폭 요구량을 구할 수 있습니다.

$$per\_leaf\_requirement = 500 \text{ 기가비트} * 50\% \; remote\_ratio = 250 \text{ 기가비트}$$

25GbE 네트워크를 사용하는 이 예에서는 리프 스위치 간에 250기가비트의 대역폭이 필요합니다. 그러나 이 대역폭은 스파인 스위치 사이에 분산되므로 추가 계산이 필요합니다.

각 리프 스위치에서 각 스파인 스위치로의 업스트림 요구량을 구하려면 스파인 스위치 간에 원격 부하가 균형을 유지하므로 원격 대역폭 요구량을 스파인 스위치 수로 나눕니다.

$$per\_leaf\_to\_spine\_requirement = \frac{per\_leaf\_requirement}{number\_of\_spine\_switches}$$

이 예에서 각 리프 스위치는 스파인 스위치의 메시를 통해 250기가비트의 원격 대역폭을 요구할 것으로 예상됩니다. 이 부하는 스파인 스위치(2개로 가정) 사이에 분산되므로 각 리프와 스파인 사이의 총 대역폭은 다음과 같이 계산됩니다.

$$per\_leaf\_to\_spine\_requirement = \frac{250 \text{ 기가비트}}{2 \; spine \; switches} = 125 \; \frac{\text{기가비트}}{spine \; switch}$$

따라서 비차단 토폴로지의 경우 총 200기가비트 연결을 위한 2개의 100기가바이트 연결이 각 리프 및 스파인 스위치 사이에 충분한 대역폭을 제공합니다. 또는 40기가비트 연결 4개로 125Gb/s를 나눌 수 있습니다.

각 리프 스위치에서 각 스파인 스위치로의 필요한 대역폭의 양을 결정하는 공식은 다음과 같이 요약할 수 있습니다.

$$\frac{downstream\_bandwidth\_requirement * ((number\_of\_racks - 1) \; / number\_of\_racks)}{number\_of\_spine\_switches}$$

참고: 복제가 구현된 시스템에서는 이러한 계산에 추가 백엔드 복제 스토리지 트래픽을 포함해야 합니다. 그러면 이러한 예에서의 요구량이 두 배로 늘어날 수 있습니다(예: 리프 스위치에 4개의 25기가비트 인터페이스).

15.7 FHRP 엔진

노드에 Cisco vPC 및 IP 수준 이중화를 사용하는 라우팅 액세스 아키텍처의 경우 노드 기본 게이트웨이에 FHRP를 사용하는 것이 좋습니다. 이렇게 하면 리프 스위치 장애가 발생하는 경우 기본 게이트웨이가 다른 리프 스위치로 페일오버할 수 있습니다. FHRP 엔진은 사용하는 스위치 공급업체에 따라 달라질 수 있습니다. Cisco 아키텍처를 사용하는 경우 HSRP가 사용됩니다. Dell 스위치의 경우 VRRP가 사용됩니다.

Aggregation Switch 1

```
interface Vlan103
no shutdown
mtu 9216
no ip redirects
ip address 192.168.103.2/24
no ipv6 redirects
hsrp version 2
hsrp 103
authentication text <text>
preempt
priority <value>
ip 192.168.103.1
```

Aggregation Switch 2

```
interface Vlan103
no shutdown
mtu 9216
no ip redirects
ip address 192.168.103.3/24
no ipv6 redirects
hsrp version 2
hsrp 103
authentication text <text>
preempt
ip 192.168.103.1
```

그림 15 Cisco Nexus 집선 스위치의 한 쌍에 대한 FHRP 엔진 구성 예. 활성 vPC 피어는 FHRP의 기본 역할을 해야 하며 백업 vPC 피어는 FHRP 보조 역할을 해야 합니다.

16 VMware 고려 사항

네트워크 연결이 ESXi에서 가상화되지만, 이 문서에 설명된 동일한 물리적 네트워크 레이아웃 원칙이 적용됩니다. Dell EMC PowerFlex 담당자와의 논의를 거치지 않았다면 특히 MDM 트래픽을 전달하는 링크에 MLAG를 사용해서는 안 됩니다.

MDM 또는 SDS를 실행하는 가상 머신의 네트워크 스택 또는 VMkernel SDC에서 사용하는 네트워크 스택의 관점에서 물리적 네트워크를 생각하면 도움이 됩니다. 게스트 또는 호스트 수준 네트워크 스택의 요구 사항을 고려하여 물리적 네트워크에 적용하면 가상 스위치 레이아웃에 대해 정보 기반의 의사 결정을 내릴 수 있습니다.

참고: 버전 3.5의 기본 비동기식 복제는 아직 VMware 기반 하이퍼 컨버지드 시스템에서 지원되지 않습니다. 따라서 Linux 기반 시스템에 대해 위에서 언급한 IP 및 처리량 고려 사항은 이 경우에 바로 적용되지 않습니다. 그러나 사용자가 앞으로의 계획을 수립하려는 경우 섹션 7.2.3에 설명된 추가 처리량 고려 사항을 고려해야 합니다.

16.1 IP 수준 이중화

이중 서브넷 구성을 사용하여 네트워크 링크 이중화가 제공되면 별도의 가상 스위치가 두 개 필요합니다. 가상 스위치마다 고유한 물리적 업링크 포트가 있으므로 필요합니다. PowerFlex가 하이퍼 컨버지드 모드에서 실행되는 경우 이 구성에는 SDC용 VMkernel, SDS용 VM 네트워크, 물리적 네트워크 액세스용 업링크의 3가지 인터페이스가 있습니다. PowerFlex는 기본적으로 이 모드에서의 설치를 지원합니다.

16.2 LAG 및 MLAG

LAG 또는 MLAG를 사용하는 경우에는 분산형 가상 스위치의 사용이 필요합니다. 표준 가상 스위치는 LACP를 지원하지 않으므로 권장되지 않습니다. LAG 또는 MLAG를 사용하는 경우 물리적 업링크 포트에서 본딩이 이루어집니다.

vSphere 플러그인을 사용한 PowerFlex 설치에서는 LAG 또는 MLAG 설치를 기본 지원하지 않습니다. 대신 PowerFlex 구축 전에 생성해 설치 과정에서 이를 선택할 수 있습니다.

SDS 또는 SDC를 실행하는 노드에 스위치에 대한 집선 링크가 있는 경우 "소스 및 대상 IP 주소" 또는 "소스 및 대상 IP 주소 및 TCP/UDP 포트"를 사용하도록 물리적 업링크 포트의 해시 모드를 구성해야 합니다.

이러한 방법은 원하는 경우 두 번째 수준의 이중화로만 사용하는 것이 좋습니다.

16.3 SDC

SDC는 PowerFlex 스토리지 클라이언트를 구현하는 ESXi용 커널 드라이버입니다. 이는 ESXi 커널에서 실행되기 때문에 다른 PowerFlex 구성 요소와의 통신에 하나 이상의 VMkernel 포트를 사용합니다. 여기서는 기본 IP 수준 이중화를 구현하기 위한 일반적인 권장 사항을 반복합니다. 이 경우 각 VMkernel 포트가 별개의 물리적 포트에 매핑됩니다. 두 번째 수준의 이중화가 필요한 경우 IP 수준 이중화 외에도 분산 스위치 레이어에 LAG 또는 MLAG를 구현할 수 있습니다.

16.4 SDS

SDS는 ESXi에서 가상 스토리지 어플라이언스(SVM)의 일부로 구축됩니다. 다시 말하지만, 권장되는 구현 방식에서는 기본 IP 수준 이중화를 사용하며, 이때 각 서브넷은 자체 가상 스위치 및 물리적 업링크 포트에 할당됩니다. 두 번째 수준의 이중화가 필요한 경우 IP 수준 이중화 외에도 분산 스위치 레이어에 LAG 또는 MLAG를 구현할 수 있습니다.

16.5 MDM

MDM은 ESXi에서 가상 스토리지 어플라이언스(SVM)의 일부로 구축됩니다. IP 수준 이중화를 사용하는 것이 좋습니다. **따라서 단일 MDM은 2개 이상의 개별 가상 스위치를 사용해야 합니다.**

17 가상 및 소프트웨어 정의 네트워킹

향후 업데이트에 대해 설명할 내용이 많지만 SDN 지원에 대한 일반적인 오해를 해소하기 위해 간략한 설명으로 알려드리겠습니다.

17.1 Cisco ACI

Dell Technologies는 Cisco ACI 기반 PowerFlex에 대해 직접적인 지원 또는 완전한 지원을 제공하지 않습니다. 특히 Cisco ACI를 통한 백엔드 스토리지 트래픽을 지원하지 않습니다. 단 프런트엔드 고객 트래픽이 ACI 패브릭을 통해 흐르는 듀얼 네트워크 확장을 통해서는 지원 가능합니다.

17.2 Cisco NX-OS

Dell Technologies는 NX-OS 독립 실행형 소프트웨어를 사용하는 VxLAN EVPN 리프 스파인 패브릭을 지원합니다.

18 검증 방법

18.1 PowerFlex 기본 툴

네트워크 성능을 모니터링하는 두 가지 기본 내장 툴은 다음과 같습니다.

1. SDS 네트워크 테스트
2. SDS 네트워크 레이턴시 측정 테스트

18.1.1 SDS 네트워크 테스트

SDS 네트워크 테스트 "`start_sds_network_test`"의 사용은 Dell EMC PowerFlex v3.5 CLI Reference Guide에서 다룹니다. 실행 후 결과를 가져오려면 "`query_sds_network_test_results`" 명령을 사용합니다.

`parallel_messages` 및 `network_test_size_gb` 옵션은 테스트를 실행하는 링크의 최대 네트워크 대역폭보다 2배 이상 크게 설정해야 합니다. 예: 단일 10GbE NIC = 1,250메가바이트 * 2 = 2,500메가바이트 또는 올림하여 3기가비트. 이 경우 명령에 "`--network_test_size_gb 3`" 매개변수를 사용해야 합니다. 이렇게 하면 네트워크에 충분한 대역폭이 전송되어 일관된 테스트 결과를 얻을 수 있습니다. 25GbE 네트워크 구성의 경우 단일 25GbE NIC = 3,125메가바이트 * 2 = 6,250메가바이트 또는 6기가비트입니다. 이 경우 명령에 "`--network_test_size_gb 6`"이 포함되어야 합니다.

병렬 메시지 크기는 시스템에 있는 코어의 총 개수와 같아야 하며, 최대 구성은 16입니다.

참고: 이 테스트는 구성된 SDS 네트워크별로 각 SDS에서 실행해야 합니다.

예시 출력:

```
scli --start_sds_network_test --sds_ip 10.248.0.23 --network_test_size_gb 8 --parallel_messages 8
Network testing successfully started.

scli --query_sds_network_test_results --sds_ip 10.248.0.23SDS with IP
10.248.0.23 returned information on 7 SDSs
    SDS 6bfc235100000000 10.248.0.24 bandwidth 2.4 GB (2474 MB) per-second
    SDS 6bfc235200000001 10.248.0.25 bandwidth 3.5 GB (3592 MB) per-second
    SDS 6bfc235400000003 10.248.0.26 bandwidth 2.5 GB (2592 MB) per-second
    SDS 6bfc235500000004 10.248.0.28 bandwidth 3.0 GB (3045 MB) per-second
    SDS 6bfc235600000005 10.248.0.30 bandwidth 3.2 GB (3316 MB) per-second
    SDS 6bfc235700000006 10.248.0.27 bandwidth 3.0 GB (3056 MB) per-second
    SDS 6bfc235800000007 10.248.0.29 bandwidth 2.6 GB (2617 MB) per-second
```

위 예에서는 네트워크 세그먼트의 다른 모든 SDS에 대해 테스트하여 SDS의 네트워크 성능을 확인할 수 있습니다. 초당 속도가 예상하는 네트워크 구성 성능에 가까워야 합니다.

18.1.2 SDS 네트워크 레이턴시 측정 테스트

"query_network_latency_meters" 명령을 사용하여 SDS 구성 요소 간의 평균 네트워크 레이턴시를 표시할 수 있습니다. 쓰기 성능을 제대로 발휘하려면 SDS 구성 요소 간의 레이턴시가 짧아야 합니다. 이 테스트를 실행할 때 10기가비트 이상의 네트워크 연결을 사용하는 경우 몇백 마이크로초보다 높은 잔차 또는 레이턴시를 찾아야 합니다.

참고: 이 테스트는 각 SDS *및* 각 SDS 네트워크에서 실행해야 합니다.

예시 출력:

```
scli --query_network_latency_meters --sds_ip 10.248.0.23
SDS with IP 10.248.0.23 returned information on 7 SDSs

SDS 10.248.0.24
  Average IO size: 8.0 KB (8192 Bytes)
  Average latency (micro seconds): 231

SDS 10.248.0.25
  Average IO size: 40.0 KB (40960 Bytes)
  Average latency (micro seconds): 368

SDS 10.248.0.26
  Average IO size: 38.0 KB (38912 Bytes)
  Average latency (micro seconds): 315

SDS 10.248.0.28
  Average IO size: 5.0 KB (5120 Bytes)
  Average latency (micro seconds): 250

SDS 10.248.0.30
  Average IO size: 1.0 KB (1024 Bytes)
  Average latency (micro seconds): 211

SDS 10.248.0.27
  Average IO size: 9.0 KB (9216 Bytes)
  Average latency (micro seconds): 252

SDS 10.248.0.29
  Average IO size: 66.0 KB (67584 Bytes)
  Average latency (micro seconds): 418
```

18.2 Iperf, NetPerf 및 Tracepath

참고: PowerFlex를 구성하기 전에 네트워크의 유효성을 검사하려면 Iperf 및 NetPerf를 사용해야 합니다. Iperf 또는 NetPerf에서 문제가 발견되면 조사가 필요한 네트워크 문제가 있을 수 있습니다. Iperf/NetPerf에서 문제가 발견되지 않는 경우 추가적으로 보다 정확한 검증을 위해 PowerFlex 내부 검증 툴을 사용하십시오.

Iperf는 IP 네트워크에서 가능한 최대 대역폭을 측정하는 데 사용할 수 있는 트래픽 생성 툴입니다. Iperf 기능을 사용하면 다양한 매개변수, 그리고 대역폭, 손실 및 기타 측정 사항에 대한 보고서를 튜닝할 수 있습니다. Iperf를 사용하는 경우 다중 병렬 클라이언트 스레드로 실행해야 합니다. IP 소켓당 스레드 8개를 사용하는 것이 좋습니다.

NetPerf는 다양한 네트워킹 유형의 성능을 측정하는 데 사용할 수 있는 벤치마크입니다. 이는 단방향 처리량과 전체 레이턴시에 대한 테스트를 제공합니다.

경로에 따른 MTU 크기를 검색하는 데에는 Linux "`tracepath`" 명령을 사용할 수 있습니다.

18.3 네트워크 모니터링

네트워크가 최적의 용량으로 작동하지 못하게 하는 문제를 파악하고 네트워크 성능 저하를 방지하려면 네트워크 상태를 모니터링해야 합니다. 시중에서 사용할 수 있는 여러 가지 네트워크 모니터링 도구가 있으며, 도구마다 다양한 기능을 제공합니다.

Dell Technologies는 다음 영역을 모니터링하는 것을 권장합니다.

- 입력 및 출력 트래픽
- 오류, 폐기, 오버런
- 물리적 포트 상태

18.4 네트워크 문제 해결 기본 사항

- Ping을 사용하여 SDS 및 SDC 간의 전체적인 연결을 확인합니다.
- 양방향으로 구성 요소 간의 연결을 테스트합니다.
- SDS 및 MDM 통신은 네트워크만을 왕복하는 시간이 1밀리초를 초과해서는 안 됩니다.
- Ping을 사용하여 구성 요소 간 왕복 레이턴시를 확인합니다.
- 스위치 측의 포트 오류, 폐기 및 오버런을 확인합니다.
- PowerFlex 노드가 작동하는지 확인합니다.
- 모든 노드에서 PowerFlex 프로세스가 설치 및 실행되고 있는지 확인합니다.
- 특히 점보 프레임을 사용하는 경우 모든 스위치 및 서버의 MTU를 확인합니다.
- 사이트 간 SDR 통신의 MTU가 WAN에 적합한지 확인합니다.
- 사이트 간 SDR 통신에 대한 정적 라우팅 구성을 확인하고 WAN을 통한 전체적인 연결을 테스트합니다.
- 가능한 경우 10기가비트 이더넷 대신 25기가비트 이상의 이더넷을 사용합니다.
- OS 이벤트 로그에서 NIC 오류, 높은 NIC 오버런 비율(2% 초과) 및 패킷 손실을 확인합니다.
- 유효한 NIC와 연결되지 않은 IP 주소를 확인합니다.
- PowerFlex에 필요한 네트워크 포트가 네트워크 또는 노드에 의해 차단되지 않았는지 확인합니다.
- 이벤트 로그 또는 OS 네트워크 명령을 사용하여 PowerFlex를 실행 중인 OS에서의 패킷 손실을 확인합니다.
- 노드에서 실행되는 다른 애플리케이션이 PowerFlex에 필요한 TCP 포트를 사용하려 하지 않는지 확인합니다.
- 풀 듀플렉스, 자동 협상, 네트워크에서 지원하는 최대 속도를 사용하도록 모든 NIC를 설정합니다.

- PowerFlex 기본 툴의 테스트 출력을 확인합니다.
- RAID 컨트롤러의 구성이 잘못되었는지 확인합니다(네트워크와 관련되지는 않지만 일반적인 성능 문제).
- 문제가 발생한 경우 로그가 덮어써지기 전에 최대한 빨리 로그를 수집합니다.
- 기타 문제 해결, 로그 수집 정보, FAQ는 Troubleshoot and Maintain Dell EMC PowerFlex v3.5 및 PowerFlex v3.5 Log Collection Technical Notes에서 확인할 수 있습니다.

19 결론

선택한 구축 옵션, 네트워크 토폴로지, 성능 요구 사항, 이더넷, 동적 IP 라우팅 및 검증 방법이 모두 강력하면서 환경 친화적이고 지속 가능한 네트워크 설계의 바탕이 됩니다. Dell EMC PowerFlex 클러스터는 다양한 노드 유형, 스토리지 미디어 및 구축 구성이 포함된 1,024개의 노드로 스케일 업할 수 있으므로 향후 확장을 위해 네트워크 설치 규모를 조정해야 합니다. PowerFlex를 컴퓨팅 및 스토리지가 동일한 노드 세트에 상주하는 하이퍼 컨버지드 모드나 스토리지 및 컴퓨팅 리소스가 분리된 2계층 모드로 구축할 수 있다는 점도 의사 결정에 영향을 미칩니다. 탁월한 성능, 확장성 및 유연성을 실현하려면 비즈니스 요구 사항을 고려하여 네트워크를 설계해야 합니다. 이 가이드의 원칙과 권장 사항을 따르면 회복탄력성과 확장성이 뛰어난 고성능 블록 스토리지 인프라스트럭처를 구축할 수 있습니다.