DELL Technologies

最佳实践

Dell EMC PowerFlex：网络最佳实践和设计注意事项

PowerFlex 版本 3.5.x

摘要

本文档介绍 Dell EMC PowerFlex™ 软件定义的存储的核心概念，以及为 PowerFlex 系统的网络（包括带复制的单站点和多站点部署）进行设计、故障处理和维护的最佳实践。

2021 年 4 月

H18390.3

修订记录

日期	描述
2021 年 4 月	有关虚拟网络和动态路由的更新
2021 年 1 月	增加了包容性语言除外责任
2020 年 6 月	PowerFlex 3.5 版本和品牌重塑 — 复制的重写和更新
2019 年 5 月	VxFlex OS 3.0 版本 — 添加和更新
2018 年 7 月	VxFlex OS 品牌重塑和整个改写 — 添加 VXLAN
2016 年 6 月	增加 LAG 覆盖范围
2015 年 11 月	初始文档

致谢

内容所有者：Brian Dean，存储技术营销

支持：Neil Gerren、Igal Moshkovich、Matt Hobbs、Dan Aharoni、Rivka Matosevich

目录

执行摘要

Dell EMC™ PowerFlex™ 产品系列采用 PowerFlex 软件定义的存储，这是一个横向扩展块存储服务，旨在提供灵活性、弹性和简易性，同时大规模实现可预测的高性能和抗风险能力。PowerFlex 存储软件以前称为 VxFlex OS，支持各种部署选项，具有多种操作系统和虚拟机管理程序功能。

PowerFlex 系列目前包括一个机架级和两个节点级产品：设备和就绪型节点。本文档主要侧重于存储虚拟化软件层本身，主要与就绪型节点相关，但希望了解基于 PowerFlex 的成功存储系统所需网络的任何人都会对此感兴趣。

PowerFlex 机架是面向现代数据中心且设计完善的机架级系统。在机架解决方案中，网络经过预先配置和优化，并由 PowerFlex Manager (PFxM) 指定、实施和维护设计。本文档不涉及机架部署情况。对于其他 PowerFlex 系列解决方案，必须设计和实施一个适当的网络。从 PFxM 3.6 版本开始，设备允许使用不受支持的商用级交换机，只要它们符合特定标准，并配置为与 PFxM 部署的拓扑匹配。我们将在下面介绍这方面的内容。

成功的 PowerFlex 部署取决于适当设计的网络拓扑。本文档提供有关网络选择的指导，还介绍了这些选择如何与不同 PowerFlex 组件的流量类型相关。本文档介绍了软件版本 3.5 中引入的各种场景，包括使用 PowerFlex 原生异步复制的超融合注意事项和部署。本文档还介绍了一般以太网注意事项、网络性能、动态 IP 路由、网络虚拟化、VMware® 环境内的实施、验证方法和监控建议。

受众和使用

本文档面向 IT 管理员、存储架构师和 Dell Technologies™ 合作伙伴和员工。本文档旨在让那些不是网络专家的读者也能读懂。但本文档假定读者对 IP 网络有适当的了解。

熟悉 PowerFlex (VxFlex OS) 的读者可以选择跳过“PowerFlex 功能概述”和“PowerFlex 软件组件”部分的许多内容。但应注意新的 Storage Data Replicator (SDR) 组件。

本指南提供了一组基本的网络最佳实践。本文档并未涵盖 PowerFlex 的每一种网络最佳实践或配置。PowerFlex 技术专家可能会推荐比本指南更全面的最佳实践。

本文档中的示例通常会使用 Cisco Nexus® 交换机，但同样的原则一般适用于任何网络供应商。[1] 为了方便起见，我们通常将运行至少一个 PowerFlex 软件组件的任何服务器简称为 PowerFlex 节点，而不区分消费选项。

在本文档末尾的“建议摘要”部分再次以**粗体**列出了贯穿始终的具体建议。

---

[1] 有关使用戴尔网络设备的一些指导，请参阅白皮书《VxFlex Network Deployment Guide using Dell EMC Networking 25GbE switches and OS10EE》。

1 PowerFlex 功能概述

PowerFlex 是一个存储虚拟化软件，可从直连存储创建基于服务器和 IP 的 SAN，以便按需提供灵活、可扩展的性能和容量。作为传统 SAN 基础架构的替代方案，PowerFlex 结合了各种存储介质，以便创建具有不同性能和数据服务选项的块存储虚拟池。PowerFlex 提供企业级数据保护、多租户功能和企业功能，例如线内压缩、QoS、精简资源调配、快照和原生异步复制。PowerFlex 提供以下好处：

出色的可扩展性 — PowerFlex 可以从仅有几个节点开始，然后在群集中扩展到几百个节点。添加设备或节点时，PowerFlex 会均匀地自动重新分发数据，从而确保实现完全平衡的分布式存储池。

卓越性能 — PowerFlex 存储池中的每个存储介质设备都用于处理 I/O 操作。这种大规模的资源 I/O 并行处理可消除瓶颈。吞吐量和 IOPS 与添加到存储池的存储设备数成正比进行扩展。自动进行性能和数据保护优化。

难以抗拒的经济性 — PowerFlex 不需要光纤通道结构或专用组件（例如，HBA）。过时的硬件无需断代升级。只需从系统中移除发生故障或过时的组件，同时添加新组件并使数据重新平衡。通过这种方式，与传统 SAN 相比，PowerFlex 可以降低存储解决方案的成本和复杂性。

出色的灵活性 — PowerFlex 提供灵活的部署选项。在双层部署中，应用程序和存储软件安装在单独的服务器池中。使用双层部署，让计算和存储团队可以保持运营自主权。在超融合部署中，应用程序和存储安装在一个共享服务器池中，占用空间少且成本低。在扩展计算和存储资源时，还可以混合这些部署模式，提供出色的灵活性。

超高弹性 — 存储和计算资源可以根据需要增加或减少。系统动态地自动重新平衡数据。可以按小增量或大增量添加和移除。无需进行容量规划或复杂的重新配置。计划外组件丢失会触发重建操作，以保持数据保护。添加组件会触发重新平衡，提高可用性能和增加容量。无需操作员干预即可在后台自动进行重建和重新平衡操作，并且不会对应用程序和用户造成停机。

适合企业和服务提供商的基本功能 — 利用服务质量控制可实现动态地管理资源使用情况，限制所选客户端可以使用的性能（IOPS 或带宽）。PowerFlex 可为数据备份和克隆提供即时的可写快照。操作员可以使用两种不同的数据布局之一创建池，确保为工作负载提供更好的环境。而且可以在需求变化时在不同池之间进行实时和无中断的迁移。精简资源调配和线内数据压缩可实现存储节约和高效的容量管理。对于版本 3.5，PowerFlex 提供原生异步复制，用于灾难恢复、数据迁移、测试场景和工作负载分流。

PowerFlex 通过保护域和存储池提供多租户功能。保护域让用户可以隔离特定的节点和数据集。存储池可用于进一步的数据隔离、分层和性能管理。例如，对性能要求高的业务关键型应用程序和数据库的数据可以存储在高性能 SSD、NVMe 或基于 SCM 的存储池中，以实现更低的延迟，而访问频率较低的数据可以存储在使用低成本、高容量 SSD 构建且每天驱动器写入较少的池中。此外，卷可以实时地从一个池迁移到另一个池，而不会中断工作负载。

2 PowerFlex 软件组件

PowerFlex 主要由三个软件组件组成：Storage Data Server (SDS)、Storage Data Client (SDC) 和 Meta Data Manager (MDM)。3.5 版引入了一个支持复制的新组件 — Storage Data Replicator (SDR)。

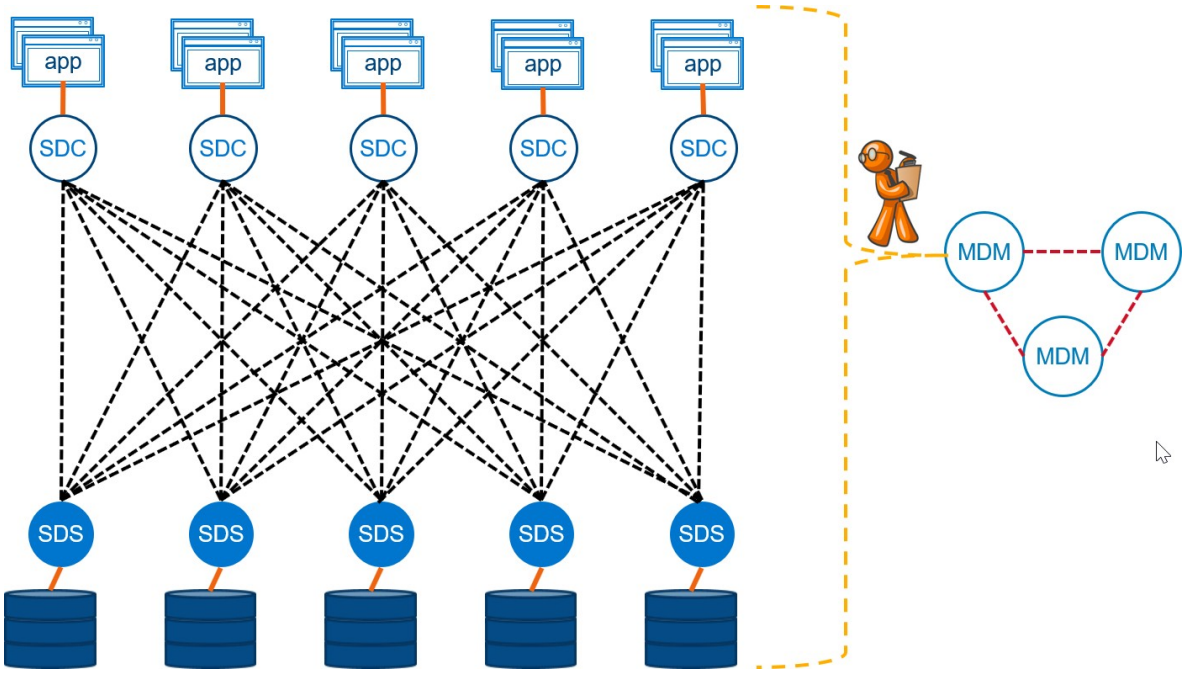

图 1 PowerFlex 部署的逻辑图解。SDC 可用的每个卷分布在许多运行 SDS 的系统中，每个 SDC 都有冗余路径连接到为卷提供服务的每个 SDS。Meta Data Manager (MDM) 群集驻留在它们监控系统时所在的数据路径之外，协调数据布局并更新 SDC（如果发生任何更改）。

2.1 Storage Data Server (SDS)

Storage Data Server (SDS) 是一个用户空间服务，将节点和服务器中的原始本地存储聚合在一起，并作为 PowerFlex 群集的一部分提供。SDS 是服务器端软件组件。参与为其他节点提供数据的任何服务器都安装了 SDS 服务并在其上运行。一组 SDS 构成了 PowerFlex 持久层。

SDS 协同工作，维护用户数据的冗余副本，保护彼此免受硬件损失，并在硬件组件出现故障时重建数据保护。SDS 可以使用 SSD、基于 PCIe 的闪存、存储级内存、旋转磁盘介质、可用 RAM 或它们的任意组合。

SDS 原生可以在各种 Linux 版本或 ESXi 上的虚拟设备中运行。一个 PowerFlex 群集最多可有 512 个 SDS。

SDS 组件之间可以直接通信，而 SDS 集合完全成网状。SDS 针对重建、重新平衡和 I/O 并行进行了优化。SDS 组件之间的用户数据布局通过**存储池、保护域和故障集**进行管理。

SDC 使用的客户端卷放置在**存储池**内。存储池用于以驱动器级别的粒度在逻辑上聚合类似类型的存储介质。存储池根据容量和性能提供不同级别的存储服务。

通过**保护域**以节点级粒度进行管理，防止节点、设备和网络连接出现故障。保护域是 SDS 组，其中维护了用户数据副本。

利用**故障集**防止冗余副本驻留在可能会一起出现故障的一组节点中（例如，整个机架），从而使非常大的系统可以容忍多个节点同时发生故障。

2.2 Storage Data Client (SDC)

Storage Data Client (SDC) 允许操作系统或虚拟机管理程序访问由 PowerFlex 群集提供的数据。SDC 是一个客户端软件组件，原生可以在 Windows®、各种 Linux、IBM AIX®、ESXi® 和其他操作系统上运行。它与软件 HBA 类似，但经过优化，可并行使用多个网络路径和端点。

SDC 为运行它的操作系统或虚拟机管理程序提供对逻辑块设备（称为“卷”）的访问。卷类似于传统 SAN 中的 LUN。每个逻辑块设备为数据库或文件系统提供原始存储，并在客户端节点上显示为本地设备。

根据数据块在卷中的位置，SDC 知道要联系哪个 Storage Data Server (SDS) 端点。SDC 从运行 PowerFlex 的其他系统直接使用分布式存储资源。SDC 不与其他 SDC 共享单个协议目标或网络端点。SDC 均匀和自主地分发负载。

SDC 非常轻便。SDC 至 SDS 通信本质上是参与存储池的所有 SDS 存储服务器的多路径通信。这与 iSCSI（多个客户端以单个协议端点为目标）等方法有着明显的不同。SDC 通信的广泛分布特征可实现更好的性能和可扩展性。

SDC 允许共享卷访问用于群集等用途。SDC 不需要 iSCSI 启动器、光纤通道启动器或 FCoE 启动器。SDC 为简易性、速度和效率而优化。一个 PowerFlex 群集最多可有 1024 个 SDC。

2.3 Meta Data Manager (MDM)

MDM 可控制 PowerFlex 系统的行为。它们确定并发布客户端与其卷数据之间的映射；它们跟踪系统的状态；它们还会向 SDS 组件发出重建和重新平衡指令。

MDM 在 PowerFlex 中确立了仲裁的概念。它们是 PowerFlex 中唯一紧密成群聚集的组件。它们具有权威性、冗余和高可用性。在 I/O 操作或 SDS 至 SDS 操作（例如，重建和重新平衡）期间，不会查询它们。但在硬件组件出现故障时，MDM 群集会指示在几秒钟内开始自动修复操作。MDM 群集包含至少三台服务器，以维护仲裁，但可以使用五台服务器来提高可用性。在 3 或 5 节点 MDM 群集中，始终有一个主节点。可能有一个或两个辅助 MDM 以及一个或两个仲裁节点。

2.4 Storage Data Replicator (SDR)

从版本 3.5 开始，引入了一个全新的可选软件，方便 PowerFlex 群集之间的异步复制。如果未使用复制，则一般的 PowerFlex 操作不需要 Storage Data Replicator (SDR)。在源端，SDR 是 SDC 与托管卷地址空间相关部分的 SDS 之间的中间人。当复制卷时，SDC 将写入发送到 SDR，在这里拆分写入，同时写入到复制日志并转发到相关的 SDS 服务，以便提交到本地磁盘。

SDR 在间隔日志中累积写入，直到 MDM 指示要关闭的间隔。如果卷是多卷复制一致性组的一部分，则间隔关闭同时发生。应用写入折叠，并将间隔添加到传输队列以传输到目标端。

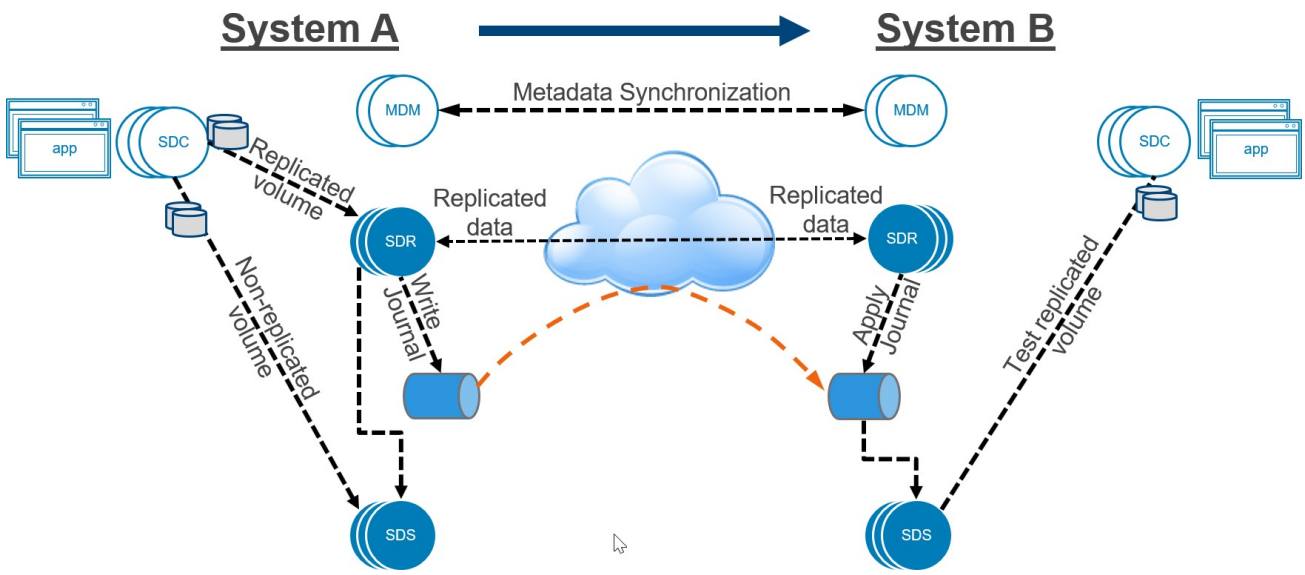

图 2　复制数据流的简化示意图。

在目标端，SDR 将数据接收到另一个日志，并将其发送到 SDS 以应用到目标副本卷。

3 流量类型

当网络体系结构反映 PowerFlex 流量模式时，PowerFlex 可获得性能、可扩展性和安全优势。在大型 PowerFlex 部署中尤其如此。组成 PowerFlex 的软件组件（SDC、SDS、MDM 和 SDR）以可预测的方式相互交流。**设计 PowerFlex 部署的架构师应了解这些流量模式，以便作出有关网络布局的明智选择。**

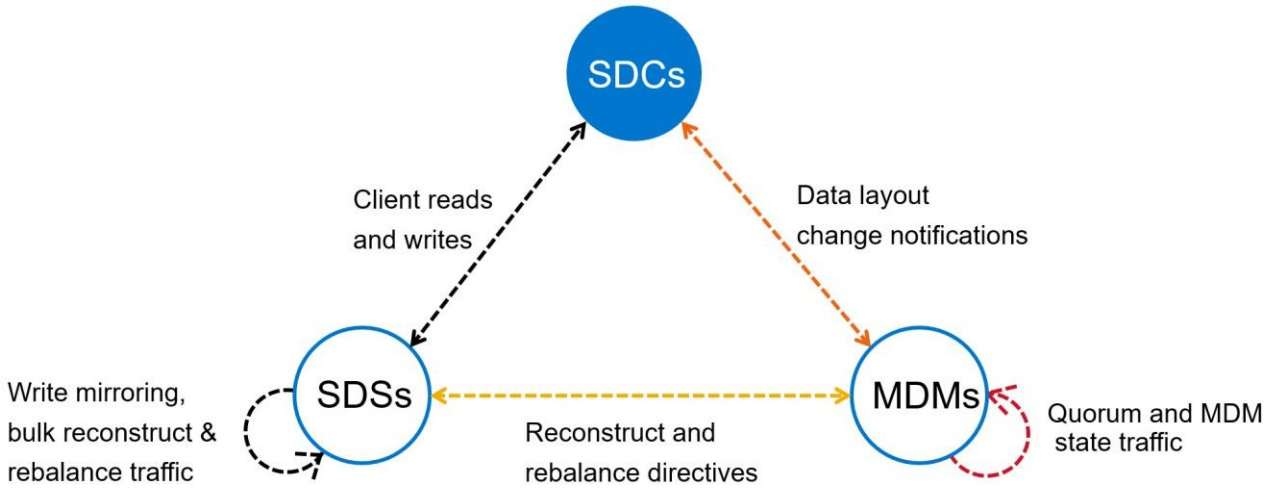

图 3 有关基本 PowerFlex 软件组件如何通信的简化图解。PowerFlex 系统会有许多 SDC、SDS 和 MDM。此图解对 SDC、SDS 和 MDM 进行了分组。SDS 和 MDM 指回自身的箭头表示与其他 SDS 和 MDM 的通信。请注意，没有 SDC 至 SDC 通信。无论 SDC、SDS 或 MDM 的物理位置如何，流量模式都相同。

在下面的讨论中，我们将前端流量与后端流量区分开来。这是逻辑上的区分，不需要物理上不同的网络。PowerFlex 允许在同一个物理网络上运行前端和后端通信，或在不同的网络上把它们分开。虽然分开不是必需的，但分开存储网络的前端和后端流量通常是优选选项。

例如，可能出于运营原因而需要进行此类分离，在这种情况下将由单独的团队管理基础架构的不同部分。然而，分离后端通信的更常见原因是为了改进重建和重新平衡性能。这也会隔离前端通信，避免网络上的争用，并减少重建/重新平衡操作期间对客户端或应用程序流量的延迟影响。

3.1 Storage Data Client (SDC) 至 Storage Data Server (SDS)

大部分前端存储流量是 SDC 和 SDS 之间的流量。前端存储流量包括到达客户端或源自客户端的所有读写流量。此网络具有高吞吐量要求。

3.2 Storage Data Server (SDS) 至 Storage Data Server (SDS)

大部分后端存储流量是 SDS 之间的流量。后端存储流量包括在 SDS 之间镜像的写入、重新平衡流量、重建流量和卷迁移流量。此网络具有高吞吐量要求。

3.3 Meta Data Manager (MDM) 至 Meta Data Manager (MDM)

MDM 用于协调群集内的运营。它们向 PowerFlex 发出指令，以便重新平衡、重建和重定向流量。它们还会协调复制一致性组、确定复制日志间隔关闭，以及与 PowerFlex 副本对等系统保持元数据同步。MDM 是冗余的，必须持续相互通信，以建立仲裁并保持对数据布局的共同理解。

MDM 不承载或直接干扰 I/O 流量。在它们之间交换的数据相对轻量级，MDM 没有对 SDS 或 SDC 流量要求相同的吞吐量。但是，MDM 的仲裁交换时间非常短 (<400 ms)，每 100 ms 进行一次。

MDM 至 MDM 流量需要稳定、可靠、低延迟的网络。MDM 至 MDM 流量被视为后端存储流量。PowerFlex 支持使用一个或多个专用于 MDM 之间流量的网络。对于生产环境，每个 MDM 至少应使用两个 10 GbE 链路，但 25 GbE 更为常见。

PowerFlex 3.5 在副本对等系统之间引入了跨群集 MDM 至 MDM 流量。这些 MDM 必须进行通信，以便控制复制流和日志状态。它们会在源和目标站点之间同步合并的复制状态。MDM 至 MDM 对等元数据同步应在延迟小于 200 ms 的 WAN 上进行。

3.4 Meta Data Manager (MDM) 至 Storage Data Client (SDC)

如果数据布局发生更改，主 MDM 必须与 SDC 通信。这可能是由于为 SDC 添加或删除了托管 SDC 卷存储的 SDS，或者将其置于维护模式或使其离线。如果将卷放入复制一致性组中，也可能会发生这种情况。主 MDM 与 SDC 之间的通信是惰性和异步的，但仍需要可靠、低延迟的网络。MDM 至 SDC 流量被视为前端存储流量。

3.5 Meta Data Manager (MDM) 至 Storage Data Server (SDS)

主 MDM 必须与 SDS 通信，以便监控 SDS 和设备运行状况，以及发出重新平衡和重建指令。MDM 至 SDS 流量需要可靠、低延迟的网络。MDM 至 SDS 流量被视为后端存储流量。

3.6 Storage Data Client (SDC) 至 Storage Data Replicator (SDR)

在复制卷的情况下，通过 SDR 路由正常的 SDC 至 SDS 流量。如果卷放入复制一致性组，则 MDM 会调整向 SDC 呈现的卷映射，并指示 SDC 向 SDR 发出 I/O 操作，然后将其传递到相关的 SDS。在 SDC 看来，SDR 就像是另一个 SDS。SDC 至 SDR 流量有很高的吞吐量要求，并且需要可靠、低延迟的网络。SDC 至 SDR 流量被视为前端存储流量。

3.7 Storage Data Replicator (SDR) 至 Storage Data Server (SDS)

当复制卷且 I/O 从 SDC 发送至 SDR 时，源系统上有两个从 SDR 至 SDS 的后续 I/O。首先，SDR 将卷 I/O 传递给关联的 SDS 以进行处理（例如压缩）并提交到磁盘。其次，SDR 将写入内容应用于日志卷。由于日志卷只是 PowerFlex 系统中的另一个卷，因此，SDR 将 I/O 发送至 SDS，这些 SDS 的磁盘组成日志卷所在的存储池。

在目标系统上，SDR 将收到的一致日志应用到支持副本卷的 SDS。在每种情况下，SDR 表现得好像它是一台 SDC。尽管如此，SDR 至 SDS 流量被视为后端存储流量。SDR 至 SDS 流量吞吐量可能很高，与要复制的卷数成正比。它需要可靠、低延迟的网络。

3.8 Metadata Manager (MDM) 至 Storage Data Replicator (SDR)

MDM 必须与 SDR 通信，以便发出日志间隔关闭、收集和报告 RPO 合规性，以及在目标卷上保持一致性。使用从对等系统传输的复制状态，MDM 命令其本地 SDR 执行日志操作。

3.9 Storage Data Replicator (SDR) 至 Storage Data Replicator (SDR)

源内或目标 PowerFlex 群集内的 SDR 不相互通信。但源系统中的 SDR 将与副本目标系统中的 SDR 通信。SDR 通过 LAN 或 WAN 网络将日志间隔发送到目标 SDR。在 SDR → SDR 流量中，延迟不那么敏感，但往返时间不得大于 200 ms。

3.10 其他流量

PowerFlex 群集中还有许多其他类型的低容量流量。其他流量包括不经常发生的管理、安装和报告。这还包括发送至 PowerFlex 网关（REST API 网关、Installation Manager 和 SNMP 陷阱发送器）、vSphere 插件程序、PowerFlex Manager 的流量、进出 Light Installation Agent (LIA) 的流量以及发送至 MDM 的报告或管理流量（用于报告的系统日志和用于管理员身份验证的 LDAP）。还包括 MDM、SDS 和 SDC 中的 CHAP 身份验证流量。有关详情，请参阅 PowerFlex 技术资源中心的《Getting to Know Dell EMC PowerFlex》指南。

SDC 不与其他 SDC 通信。这可以使用专用 VLAN 和网络防火墙来实施。

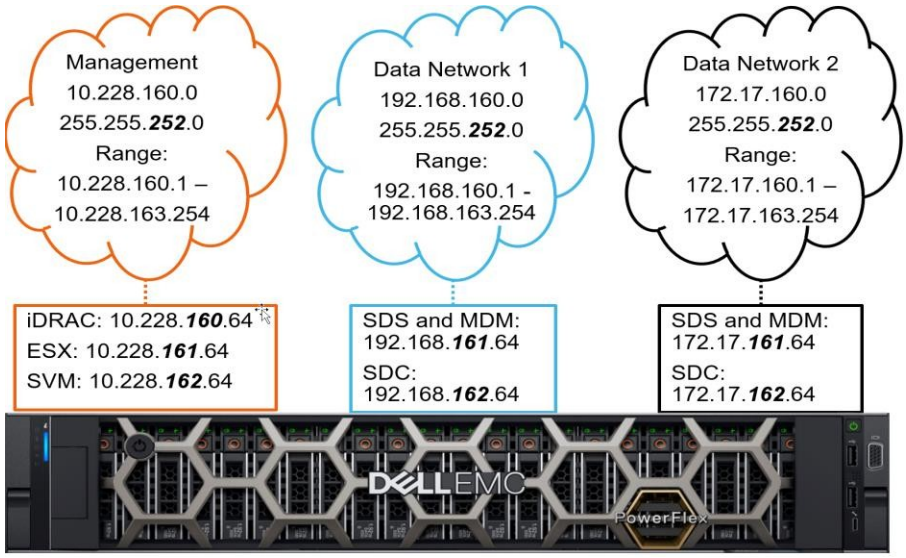

图 4　简单的 PowerFlex 超融合网络布局。按规定路线安排管理网络，并提供对 iDRAC、ESX 和存储虚拟机 (SVM) 的访问权限。冗余网络承载 SDS、MDM 和 SDC 流量。SDS 和 MDM 流量使用同一组 IP 地址。流量不像大型部署那样划分为前端（SDS、SDC、MDM）和后端（SDS、MDM）流量。192.168.160.X 和 172.17.160.X 地址空间可用于 MDM 虚拟 IP。

4 PowerFlex TCP 端口使用情况

PowerFlex 在以太网结构上运行。虽然许多 PowerFlex 协议是专有的，但所有通信都使用标准 TCP/IP 传输。

下图简要概括了 PowerFlex 软件组件之间的端口使用和通信。有些端口是固定的，不可更改，而其他一些端口可配置，可以重新分配到另一个端口。要获得完整的列表和分类，请参阅《Dell EMC PowerFlex Security Configuration Guide》中的“Port usage and change default ports”部分。

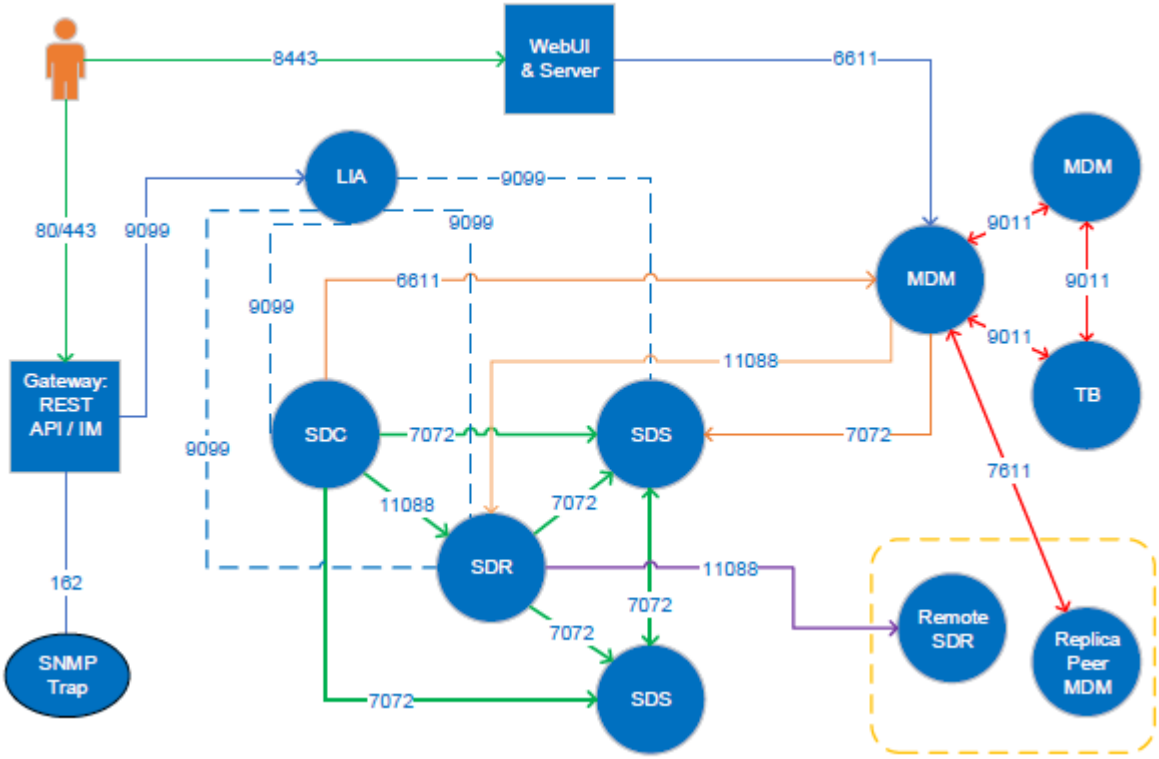

图 5 PowerFlex 软件定义的存储组件内的 TCP 端口使用情况和通信。图中的箭头指示连接发起的方向。即，箭头指向侦听服务端口。数据可能会在启动后在连接上双向传输。虚线表示在已安装组件之间的节点内部进行通信。

MDM 上的端口 25620 和 25600 以及 SDS 上的 25640 也可能会监听。这些仅由 PowerFlex 内部调试工具使用，不属于日常运营和流量的一部分。

5 网络容错

PowerFlex 组件（MDM、SDS、SDC、SDR）之间的通信应分配到不同物理网络上的至少两个子网。每个组件的 PowerFlex 网络层跨分配的多个子网提供原生链路容错和多路径。由此产生的优势是：

1. 在链路出现故障时，PowerFlex 几乎可以立即了解问题，并调整以适应带宽损失。
2. 如果使用了基于交换机的链路聚合，PowerFlex 就无法确定单链路丢失。
3. 当链路出现故障时，PowerFlex 会在 2-3 秒内在分配给 MDM、SDS 和 SDC 组件的子网中动态调整通信。这对于 SDS→SDS 和 SDC→SDS 连接尤其重要。
4. 这每一个组件都能够在最多八个子网之间负载均衡和聚合流量，从而降低交换机链路聚合维护的复杂性。而且，由于它由存储层本身进行管理，因此比交换机聚合更高效和更易于维护。

   提醒：在以前版本的 PowerFlex 软件中，如果发生与链路相关的故障，则可能会出现网络服务中断和 I/O 延迟（在 SDC→SDS 网络中会延迟长达 17 秒）。SDC 具有一般为 15 秒的超时，并且仅当已达到超时且已关闭失灵的插槽时，才会在另一个“良好的”插槽上重新发出 I/O。

   在版本 3.5 及更高版本中，PowerFlex 不再依赖于 I/O 超时，而是使用链路断开通知。发生链路断开事件后，在 2 秒后关闭所有相关的 TCP 连接，并且终止未收到响应的所有动态 I/O 消息，并由 SDC 重新发出 I/O。

完全支持原生网络路径负载均衡和基于交换机的链路聚合，但依赖原生网络路径负载均衡通常更简单。如果需要，可以将这些方法结合起来，通过中继创建两个数据路径网络，其中每个逻辑网络在每个节点上使用两个物理端口。

PowerFlex Manager 为设备执行此操作。它将链路聚合与原生多路径结合使用，提供分层且强大的网络容错。请参阅《Dell EMC PowerFlex Appliance Network Planning Guide》。

6 网络基础架构

分支-主干和扁平网络拓扑是当今 PowerFlex 的常用拓扑。扁平网络在小型网络中使用。在现代数据中心，分支-主干拓扑比传统的分层拓扑更受青睐。本节比较作为 PowerFlex 数据流量传输介质的扁平和分支-主干拓扑。

Dell Technologies 推荐使用无阻塞网络设计。采用无阻塞网络设计时，可以同时使用所有交换机端口，而不会阻断某些网络端口，以防止消息循环。因此，Dell Technologies 强烈建议不要在托管 PowerFlex 的网络上使用生成树协议 (STP)。为了实现更高的性能和可预测的服务质量，网络不应超额预订。

6.1 分支-主干网络拓扑

双层分支-主干拓扑可在分支交换机之间提供单个交换机跃点，并在端点之间提供大量带宽。适当大小的分支-主干拓扑消除了上行链路端口的超额预订。非常大的数据中心可以使用三层分支-主干拓扑。为了简单起见，本白皮书重点介绍两层分支-主干部署。

在分支-主干拓扑中，每个分支交换机连接到所有主干交换机。分支交换机不需要直接连接到其他分支交换机。主干交换机不需要直接连接到其他主干交换机。

在大多数情况下，Dell Technologies 建议使用主干交换机网络拓扑。这是因为：

- PowerFlex 可以在单个群集中横向扩展到数百个节点。
- 分支-主干体系结构让您未来无忧。它们简化了横向扩展部署，而无需重新设计网络。
- 通过分支-主干拓扑，可以同时使用所有网络链路。传统的分层拓扑必须采用生成树协议 (STP) 之类的技术，从而阻塞部分端口，防止出现环路。
- 由于消除了上行链路超额预订，因此大小合适的分支-主干拓扑可以提供可预测性更高的延迟。

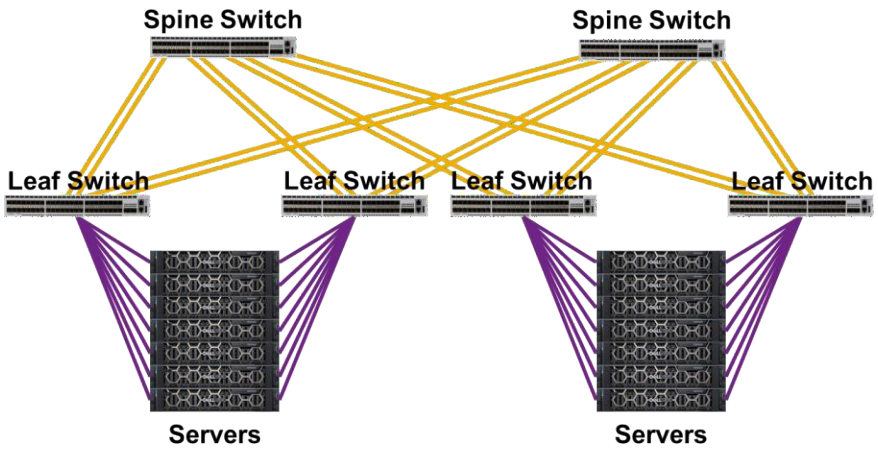

图 6 双层分支-主干网络拓扑。每台分支交换机有多条路径通往所有其他分支交换机。所有链路均处于活动状态。这提高了网络上设备之间的吞吐量。分支交换机可相互连接，以便与 MLAG 一起使用（此处未显示）。

6.2 扁平网络拓扑

扁平网络拓扑更易于实施，如果要扩展现有的扁平网络，或者网络预计不会扩展，则扁平网络拓扑可能是优选。在扁平网络中，所有交换机都用于连接主机。没有主干交换机。

但是，如果您扩展到少量接入交换机以外，则所需的额外跨链路端口可能会使扁平网络拓扑成本过高。扁平网络拓扑的应用场景包括概念验证部署和不超过几个机架的小型数据中心部署。

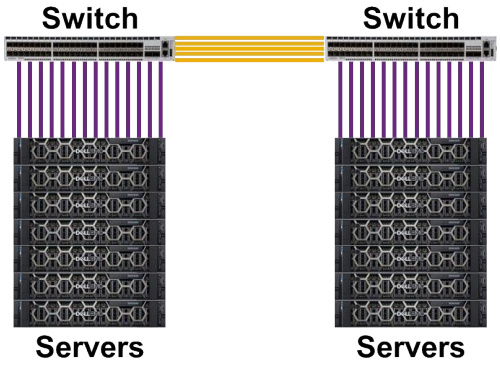

图 7 扁平网络。这一网络设计降低了成本和复杂性，但以牺牲冗余和可扩展性为代价。在这个图中，每个交换机都是单点故障。可以使用 MLAG（此处未显示）等技术构建无单点故障的扁平网络。

7 网络性能和规模调整

适当大小的网络可使网络和存储管理员不必担心单个端口或链路成为性能或运营瓶颈。使用网络管理而不是端点热点，这是 PowerFlex 的主要体系结构优势。

由于 PowerFlex 在网络中的多个点均匀地分发 I/O，因此必须相应地调整网络性能。

7.1 网络延迟

在设计网络时，网络延迟是一个重要的考虑因素。更大限度地减少网络延迟将有助于提高性能和可靠性。**为了获得更好的性能，在正常运行条件下，所有 SDS 和 SDC 通信的延迟都不应超过 1 毫秒仅网络往返时间。**由于广域网 (WAN) 的最低响应时间通常会超过此限制，因此您不应在 WAN 上运行 PowerFlex 群集。

对于一般通信、SDC、MDM 和 SDS 通信，实施异步复制的系统也不例外。在独立的 PowerFlex 群集之间复制数据，每个群集本身也应遵循低于 1 ms 的规则。不同之处在于对等系统之间的延迟。由于异步复制通常在 WAN 上发生，因此延迟要求必然会降低限制。**但是，对等 PowerFlex 群集组件之间的网络延迟（无论是 MDM←→MDM 还是 SDR←→SDR）不应超过 200 ms 的往返时间。**

应在所有组件之间双向测试延迟。这可以通过 ping 进行验证，在更普遍情况下可通过 SDS 网络延迟计量测试来验证。开放源代码工具 iPerf 可用于验证带宽。请注意，iPerf 不受 Dell Technologies 支持。本文档的“验证方法”部分详细介绍了用于验证 PowerFlex 部署的 iPerf 和其他工具。

7.2 网络吞吐量

在设计 PowerFlex 实施时，网络吞吐量是一个关键部分。在以下情况下，吞吐量很重要：减少故障节点重建所需的时间；减少数据分布不均时重新分发数据所需的时间；优化节点能够传递的 I/O 量；以及满足性能预期。

虽然 PowerFlex 软件可以部署在 1 千兆网络上以用于测试或调查目的，存储性能可能会受到网络容量的限制。**戴尔建议至少使用 10 千兆网络技术，而使用 25 千兆技术作为优选的最小链路吞吐量。**所有当前的 PowerFlex 节点都具有至少四个端口，每个端口的最低端口带宽为 25 GbE，并提供 100 GbE 端口作为有远见的选择。在考虑复制用例及其额外带宽需求时，这一点尤其重要。

此外，尽管 PowerFlex 群集本身可能是异构的，但**构成保护域的 SDS 组件应驻留在具有同等存储和网络性能的硬件上。**这是因为保护域的总带宽会受到 I/O 期间最薄弱链路的限制，并且由于所有参与组件中的卷数据发生广泛条带化，所以要执行重建/重新平衡操作。您可以把它想象成一次远足派对，行进速度不会比最慢的成员快。

混合异构操作系统和虚拟机管理程序组合时也会有类似的考虑因素。由于存在虚拟化开销，基于 VMware 的超融合基础架构具有比裸机配置更低的性能配置，并且在保护域中混合使用 HCI 和裸机节点会限制包含这两者的存储池的吞吐量，使之不超过最慢成员的性能能力。这是有可能且允许的（从存储软件的角度来看），但用户必须注意到这样带来的影响。PowerFlex 机架或设备不支持该配置。

除了吞吐量考虑因素外，**不管吞吐量要求如何，建议每个节点至少有两个独立的网络连接以实现冗余。**即使随着网络技术不断进步，这一点仍然很重要。例如，用一个 100 Gb 链路替换两个 40 Gb 链路可提高吞吐量，但这样舍弃了链路级网络冗余。

在大多数情况下，节点的网络吞吐量应匹配或超过节点上托管的存储介质的合并最大吞吐量。*换句话说，节点的网络要求与底层存储介质的总体性能成正比。*

在确定所需的网络吞吐量时，请记住，现代介质性能通常以 *MB*/s 为单位，但现代网络链路通常以 *Gb*/s 为单位。

要将 *MB*/s 转换为 *Gb*/s，首先将 *MB* 乘以 8，转换为 *Mb*，然后将 *Mb* 除以 1000，得到 *Gb*。

$$\text{Gb} = \frac{\text{MB} * 8}{1{,}000}$$

请注意，这并不完全精确，因为它没有考虑“千”的二进制定义为 1024，而 PowerFlex 中使用二进制标准，但这足以用于本文的说明目的。

7.2.1 示例：带 10 个 SSD 的仅 SDS（仅存储）节点

假定您的 1U 节点仅托管一个 SDS。这不是超融合环境，因此只需考虑存储流量。该节点包含 10 个 SAS SSD 驱动器。在理想条件下（顺序 I/O，PowerFlex 针对重建和重新平衡操作进行了优化），这每一个驱动器都可以单独提供每秒 1000 MB 的原始吞吐量。因此，底层存储介质的总吞吐量为每秒 10,000 *MB*。

$$10 * 1000\ \text{MB} = 10{,}000\ \text{MB}$$

然后使用前面给出的等式将 10,000 *MB* 转换为 *Gb*：先将 10,000 MB 乘以 8，然后除以 1000。

$$\frac{10{,}000\ \text{MB} * 8}{1{,}000} = 80\ \text{Gb}$$

在本例中，如果节点上的所有驱动器都以可能的最大速度执行读取操作，则所需的总网络吞吐量将为每秒 80 Gb。我们仅考虑读取操作，这通常足以评估网络带宽要求。单个25 或 40 Gb 链路无法提供此带宽，但从理论上来说，100 GbE 链路就足够了。但是，由于我们建议提供网络冗余，所以此节点应至少有两个 40 Gb 链路，优先选择标准的 4 个 25 GbE 配置。

提醒：仅根据组件驱动器的理论吞吐量来计算吞吐量可能会导致对单个节点的估计高得不合理。**确认节点上的 RAID 控制器或 HBA 也可以达到或超过底层存储介质的最大吞吐量。**

7.2.2 写入频繁的环境

在 PowerFlex 环境中，读和写操作产生不同的流量模式。当主机 (SDC) 发出单个 4k 读取请求时，它必须联系单个 SDS 以检索数据。4k 块从单个 SDS 传出一次。如果该主机发出单个 4k 写入请求，则必须将 4k 块传输到主 SDS，然后从主 SDS 复制到辅助 SDS。

因此，SDS 中的写入操作需要比读取操作多两倍的带宽。但是，写入操作涉及两个 SDS，而读取操作只需要一个 SDS。因此，读取与写入的带宽需求比为 1:1.5。

换句话说，对于每个 SDS，与底层存储的吞吐量相比，写入操作需要比读取操作多 1.5 倍的网络吞吐量。

在一般情况下，前面所述的存储带宽计算就足够了。**但是，如果环境中的某些 SDS 需要托管写入频繁的工作负载，请考虑增加网络容量。**

7.2.3 将卷复制到另一个系统的环境

版本 3.5 引入原生异步复制，在考虑生成的带宽时必须考虑这一点，首先是在群集内，其次是在副本对等系统之间。

7.2.3.1 复制系统内的带宽

我们在上面注意到，在复制卷时，I/O 从 SDC 发送到 SDR，之后，后续 I/O 从 SDR 发送到源系统上 SDS。SDR 首先将卷 I/O 传递给关联的 SDS 进行处理（例如，压缩）并提交到磁盘。关联的 SDS 可能不是与 SDR 位于同一节点上，并且带宽计算必须考虑到这一点。在第二步中，SDR 将传入的写入应用到日志记录卷。由于日志卷就像 PowerFlex 系统内的任何其他卷，因此，SDR 将 I/O 发送到为日志卷所在存储池提供支持的各种 SDS。*此步骤添加两个额外的 I/O，因为 SDR 先写入到为日志卷提供支持的相关主 SDS，然后主 SDS 将副本发送到辅助 SDS。*最后，SDR 会从日志卷中进行额外的读取，然后再发送到远程站点。

因此，复制卷的写入操作在源群集内所需的带宽是非复制卷的写入操作的三倍。**仔细考虑将在复制卷上运行的工作负载的写入配置文件；需要额外的网络容量来容纳额外的写入开销。**因此，在复制系统中，我们建议使用 4 个 25 GbE 或 2 个 100 GbE 网络来容纳后端存储流量。

7.2.3.2 副本对等系统之间的带宽

考虑副本对等系统之间的网络需求，我们重申，**源系统和目标系统之间的延迟不应超过 200 ms。**

日志数据在源和目标 SDR 之间传送，首先在复制配对初始化阶段传送，其次是在复制稳定状态阶段传送。应特别注意确保源和目标 SDR 之间有充足的带宽，无论是通过 LAN 还是 WAN 的带宽。使用 WAN 连接，超出可用带宽的可能性很大。虽然写入折叠可以减少发送到目标日志的数据量，但这个数据量并不总是能够轻松预测。*如果超出可用带宽，日志间隔会备份，同时增加日志卷大小和 RPO。*

按照妥善做法，我们建议要复制的所有卷的持续写入带宽不应超过总计可用 WAN 带宽的 80%。如果对等系统相互复制卷，则对等 SDR←→SDR 带宽必须同时满足两个方向的要求。如果在计算特定工作负载所需的 WAN 带宽方面需要额外的帮助，请参考并使用新版 PowerFlex Sizer。

提醒：该规模调整工具是提供给戴尔员工和合作伙伴的内部工具。如果需要 WAN 带宽规模调整协助，外部用户应咨询其技术销售专家。

7.2.3.3 复制运行状况的网络影响

虽然本白皮书的重点是 PowerFlex 网络信息最佳实践，但存储层本身的一般操作、运行状况和性能取决于已部署网络的质量和容量。这对于异步复制和日志卷的大小调整具有特殊意义。

写入峰值可能会超过建议的“0.8 * WAN 带宽”，但它们的持续时间应该很短。日志大小必须足够大，以便吸收这些写入峰值。

这一点非常重要。日志卷容量应进行调整，以适应对等系统之间的链路中断。预计一个小时的停机可能是合理的，但我们强烈建议用户规划 3 小时的停机。在中断期间，用户必须确保有足够的日志空间来处理应用程序写入。**一般而言，日志容量应按如下方式计算：峰值写入带宽 * 链路停机时间。**我们需要知道最繁忙时段的最大应用程序写入带宽。假设我们的应用程序具有 1 GB/s 的峰值写入吞吐量。3 小时为 10800 秒。因此，所需的日志容量为

$$1\ \text{GB/s} * 10800\ \text{秒} = \sim 10.55\ \text{TB}$$

但是，PowerFlex 将日志容量设置为池容量的百分比。假设我们有一个 200 TB 的存储池：

$$100 * 10.55\ \text{TB} / 200\ \text{TB} = 5.27\%$$

为了留出安全裕度，将此值四舍五入为 6%。

提醒：日志间隔中发送的卷数据不会压缩。在 PowerFlex 中，压缩用于静态数据。在精细粒度存储池中，从 SDC（用于非复制卷）或 SDR（用于复制卷）收到数据之后，在 SDS 服务中进行数据压缩。SDR 不需要知道复制对任一侧的数据布局。如果目标卷配置为压缩，则在应用日志间隔时，在目标系统 SDS 中进行压缩。

7.2.4 超融合环境

当 PowerFlex 为超融合部署时，每个物理节点运行 SDS（虚拟机管理程序上的 SDC）以及一个或多个虚拟机。从这个意义上说，超融合 PowerFlex 部署无需涉及虚拟机管理程序。超融合部署可优化硬件投资，但也会引入网络规模调整要求。

上述存储带宽计算适用于超融合环境，但也必须考虑任何虚拟机的前端带宽、虚拟机管理程序或操作系统流量以及来自 SDC 的流量。虽然虚拟机的规模调整不在本技术报告的范围之内，但这是一个优先事项。

在超融合环境中，从逻辑上将存储与其他网络流量分隔开来，这也是一个优先事项。

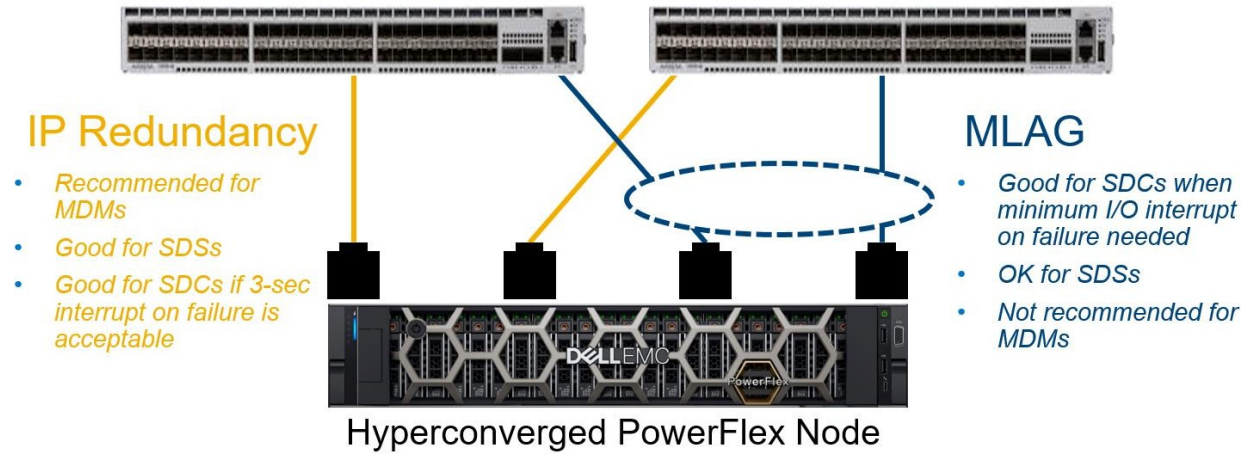

图 8 使用 4 个 25 千兆网络连接的超融合 VMware 环境示例。此主机上的 PowerFlex 流量使用端口 EthO 和 Eth1。使用原生 PowerFlex IP 多路径（而非 MLAG）提供冗余。端口 Eth2 和 Eth3 使用 MLAG 和 VLAN 标记，并提供对虚拟机管理程序和其他来宾虚拟机的网络访问权限。由于 PowerFlex 支持 VLAN 标记和链路聚合，因此可以使用其他配置。

8 网络硬件

8.1 专用 NIC

如果可能，PowerFlex 工程建议为 PowerFlex 流量使用专用网络适配器。专用网络适配器提供专用带宽和简化的故障处理。请注意，支持共享网络适配器，并且在超融合环境中可能是强制性的。

8.2 共享 NIC

虽然不是理想选择，但 PowerFlex 软件支持使用共享 NIC。如果 PowerFlex 流量将与其他非 PowerFlex 流量共享物理网络，则应实施 QoS，避免因 PowerFlex 或非 PowerFlex 流量而导致的网络拥塞或带宽不足问题。

8.3 两个 NIC 与四个 NIC 及其他配置

PowerFlex 允许通过添加额外的网络接口来扩展网络资源。**虽然不一定要分隔，但在某些情况下，分隔存储网络的前端和后端流量可能是理想的选择。**在存储和虚拟化或计算团队各自管理自己网络的双层部署中，这可能非常有用。更常见的情况是，用户会分开前端和后端网络流量，以确保与存储和应用程序相关的网络流量的性能。在所有情况下，戴尔建议使用多个接口以实现冗余、容量和速度。

PCI NIC 冗余也是一个考虑因素。**在每台服务器上使用两个双端口 PCI NIC 比使用一个四端口 PCI NIC 更可取，因为可以将两个双端口 PCI NIC 配置为在单个 NIC 发生故障时仍能继续使用。**

8.4 交换机冗余

在大多数分支-主干配置中，主干交换机和架顶式 (ToR) 分支交换机都是冗余的。这样，在 ToR 交换机出现故障时，可以继续在网络中访问机架内的组件。如果每个机架都包含一台 ToR 交换机，则 ToR 交换机故障将导致无法访问机架内的 SDS 组件。**因此，不建议使用单 ToR 交换机配置。**如果每个机架使用单台 ToR 交换机，用户应在机架级别定义故障集，以确保交换机出现故障时数据可用。

9 IP 注意事项

9.1 IPv4 和 IPv6

从版本 2.6 开始，包括 3.0 之后的所有版本，PowerFlex 在双层和超融合部署选项中提供 IPv6 支持。之前的 PowerFlex 版本仅支持 Internet Protocol version 4 (IPv4)。本白皮书中的示例专注于 IPv4。

9.2 IP 级冗余

MDM、SDS、SDR 和 SDC 可以有多个 IP 地址，因此可以驻留在多个网络中。这样就提供了负载均衡和冗余选项。
当软件组件配置为跨多个链路发送流量时，PowerFlex 可在不同的物理网络链路原生地提供冗余和负载均衡。在此配置中，为 MDM、SDR 或 SDS 提供的每个物理网络端口都将分配自己的 IP 地址，每一个地址位于不同的子网中。

使用多个子网可在网络级别提供冗余。使用多个子网还可确保在将流量从一个组件发送到另一个组件时，根据目标 IP 地址选择源组件路由表中的不同条目。因为源联系单个目标上的多个 IP 地址（每个地址对应于一个物理网络端口），所以这可防止源中的单个物理网络端口成为瓶颈。

换句话说，如果源和目标上的多个物理端口位于同一子网中，则源端口可能会出现瓶颈。例如，如果两个 SDS 共享一个子网，则每个 SDS 有两个物理端口，每个物理端口在该子网中都有自己的 IP 地址，IP 堆栈将导致源 SDS 始终选择相同的物理源端口。**由于每个端口对应于主机路由表中的不同子网，因此跨子网分隔端口可以实现负载均衡。**

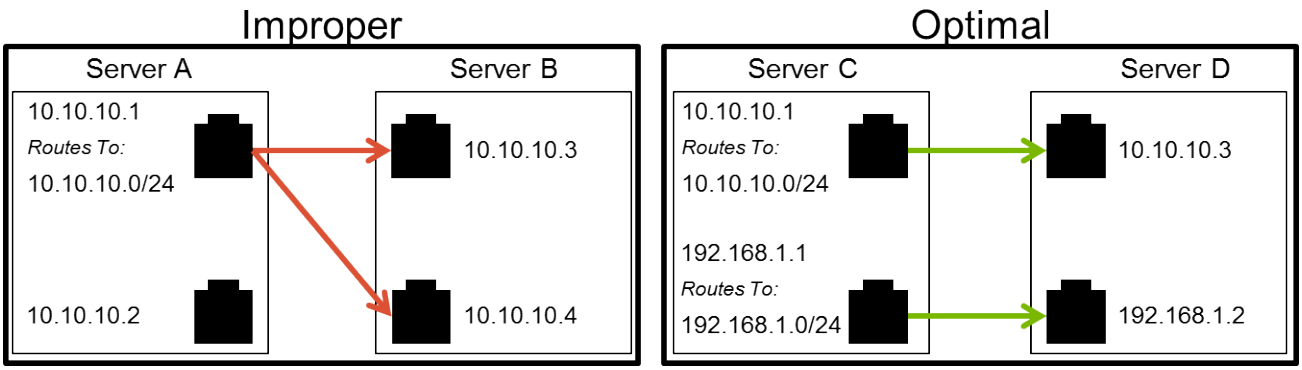

图 9 操作系统 IP 配置的比较。左侧的错误 IP 配置为所有流量使用相同的子网 (10.10.10.0/24)。当服务器 A 发起到服务器 B 的连接时，将始终为传出连接选择提供路由到 10.10.10.0/24 的网络链路。服务器 A 上的第二个网络端口不会用于传出连接。右侧的正确 IP 配置使用两个子网：10.10.10.0/24 和 192.168.1.0/24，使服务器 C 上的两个端口都可用于传出连接。提醒：在此示例中可以任意选择子网（10.10.10.0/24 和 192.168.1.0/24）：混合使用类别“A”和类别“C”只是为了视觉上的区别。

当每个 MDM 或 SDS 都可以访问多个 IP 地址时，因为 PowerFlex 了解流量模式，所以可以更有效地处理负载均衡。这样会促成性能小幅提升。
此外，链路聚合还为链路级故障切换维护它自己的一组计时器。因此，原生 PowerFlex IP 级冗余使得链路中断时的故障处理更容易。

IP 级冗余还可防止 IP 地址冲突。为防止不必要的 IP 更改或冲突，**不得在 PowerFlex MDM 或 SDC 所在的网络上部署 DHCP。**

在单独使用时，在使用 MDM 到 MDM 通信的链路中优先考虑使用 IP 级冗余，而不使用 MLAG。如果 IP 级冗余在 VLAN 中分层在冗余链路聚合组之上，那么就很好地利用了这两种技术。有关此操作的示例，请参阅《Dell EMC PowerFlex Appliance Network Planning Guide》。

10 以太网注意事项

10.1 巨型帧

PowerFlex 支持巨型帧，因此我们强烈建议为存储流量使用巨型帧。但是，根据您的网络基础架构，启用巨型帧会很具挑战性。各种网络组件对巨型帧的实施不一致，这可能会导致出现性能问题，难以进行故障处理。如果使用了巨型帧，必须在 PowerFlex 基础架构所用的每个网络组件（包括主机和交换机）和存储虚拟机（如果部署了 HCI）上启用它们。

启用巨型帧即可实现在单个以太网帧中传递更多数据。这可以减少每个节点必须处理的以太网帧总数和中断数。如果在 PowerFlex 基础架构中的每个组件上启用了巨型帧，则性能优势可提高大约 10%，具体取决于您的工作负载。

提醒：使用 PowerFlex Manager 在设备或机架系统上部署 PowerFlex 群集时，可为所有群集组件全面协调和管理节点和交换机组件上的巨型帧配置。

仔细查看网络组件，确保每个点的巨型帧配置一致。如果您不确定如何使用，我们建议在一开始禁用巨型帧。仅在拥有稳定的工作环境后才启用巨型帧，并确认您的基础架构可以支持使用巨型帧。为了确保在每条路径的所有节点上都配置了巨型帧，您可以使用实用工具（例如，Linux `tracepath` 命令）来发现路径上的 MTU 大小。在诊断巨型帧问题时，ping 也很有用。在 Linux 上，使用以下形式的命令：`ping -M do -s 8972 <ip address/hostname>`。（请注意，在这个命令中，我们从 9000 MTU 大小中减去 28 个字节的未封装数据包标头。）

有关实施巨型帧的附加信息，请参阅《PowerFlex Configure and Customize guide》。

10.2 VLAN 标记

PowerFlex 独立于服务器与接入或分支交换机之间的连接上的原生 VLAN 和 VLAN 标记。在操作系统或交换机中进行配置时，这些标记对 PowerFlex 软件是透明的。当 PowerFlex 工程进行测量时，VLAN 对性能级别没有影响。

对于 PowerFlex 设备部署，我们期望配置一组标准的统一 VLAN。请参见下面的第 19 节。

11 链路聚合组

链路聚合组 (LAG) 和多机箱链路聚合组 (MLAG) 合并端点之间的端口。端点可以是一台交换机和一台带有 LAG 的主机，也可以是两台交换机和一台带有 MLAG 的主机。链路聚合术语和实施因交换机供应商而异。Cisco Nexus 交换机上的 MLAG 功能称为虚拟端口通道 (vPC)。

LAG 使用链路聚合控制协议 (LACP) 进行设置、删除和错误处理。LACP 是一个标准，但有许多专有变体。

无论交换机供应商或托管 PowerFlex 的操作系统是什么，**在使用链路聚合组时推荐使用 LACP。不支持使用静态链路聚合。**

链路聚合可以作为 IP 级冗余的替代方案，其中每个物理端口都有自己的 IP 地址。对于某些团队来说，链接聚合的配置很简单，并且在 IP 地址耗尽是一个问题的情况下很有用。必须在运行 PowerFlex 的节点及其连接的网络设备上配置链路聚合。

无论是选择 IP 级冗余还是链路聚合，PowerFlex 都具有弹性和高性能。使用 MLAG 时，SDS 的性能接近 IP 级冗余的性能。

- **为 SDS 选择 MLAG 或 IP 级冗余应被视为一项操作决策。**
- **使用 MDM 至 MDM 流量时，强烈建议使用 IP 级冗余，而不使用 MLAG，因为 MDM 上的一个 IP 地址持续可用有助于防止故障切换，这是由于 MDM 之间的超时很短，设计为可在多个 IP 地址之间进行通信。**
- **由于在 3.5 中改进了网络故障抗风险能力，对于 SDC 组件使用的链路，通常更应该使用 IP 级冗余，而不使用 MLAG。**

11.1 LACP

LACP 会定期在聚合网络链路组中的每个物理网络链路上发送一条消息。此消息是逻辑的一部分，用于确定每个物理链路是否仍处于活动状态。网络管理员可以使用 LACP 计时器来控制这些消息的发送频率。

LACP 计时器通常可以配置为快速地（每秒一条消息）或以正常速度（每 30 秒一条消息）检测链路故障。当 LACP 计时器配置为快速地运行时，系统会快速采取纠正行动。此外，借助现代网络技术，每秒发送一条消息的相对开销很小。

在 PowerFlex SDS 和交换机之间使用链路聚合时，应将 LACP 计时器配置为快速地运行。

要建立 LACP 连接，必须将一个或两个 LACP 对等配置为使用活动模式。**因此，建议将连接到 PowerFlex 节点的交换机配置为跨链路使用活动模式。**

11.2 负载均衡

当链路聚合组中有多个网络链路处于活动状态时，端点必须选择如何在链路之间分配流量。网络管理员通过在端点上配置负载均衡方法来控制此行为。负载均衡方法通常根据源或目标 IP 地址、MAC 地址或 TCP/UDP 端口的某些组合来选择要使用的网络链路。

这种负载均衡方法称为“散列模式”。散列模式负载均衡的目的是，使进出某一对源地址和目标地址或传输端口的流量保持在同一物理链路上，前提是该链路保持活动。

散列模式负载均衡的推荐配置取决于所使用的操作系统。

如果运行 SDS 的节点已将链路聚合到交换机并运行 VMware ESX®，则应将散列模式配置为使用“源和目标 IP 地址”或“源和目标 IP 地址以及 TCP/UDP 端口”。

如果运行 SDS 的节点已将链路聚合到交换机并运行 Linux，则 Linux 上的散列模式应配置为使用“xmit_hash_policy=layer2+3”或“xmit_hash_policy=layer3+4”绑定选项。

“xmit_hash_policy=layer2+3”绑定选项使用源和目标 MAC 和 IP 地址实现负载均衡。

“xmit_hash_policy=layer3+4”绑定选项使用源和目标 IP 地址和 TCP/UDP 端口实现负载均衡。

在 Linux 上，还应使用“miimon=100”绑定选项。此选项指示 Linux 每 100 毫秒验证每个物理链路的状态。

请注意，每个绑定选项的名称可能因 Linux 发行版的不同而不同，但建议是相同的。

11.3 多机箱链路聚合组

与链路聚合组 (LAG) 类似，MLAG 提供网络链路冗余。与 LAG 不同的是，MLAG 允许单个端点（例如运行 PowerFlex 的节点）连接到多个交换机。交换机供应商在提到 MLAG 时会使用不同的名称，而 MLAG 实施通常是专有的。

PowerFlex 支持使用 MLAG，但一般不建议用于 MDM 至 MDM 流量。请参阅下一节中的注释。“负载均衡”部分介绍的选项也适用于 MLAG 的使用。

12 MDM 网络

虽然 MDM 不驻留在主机 (SDC) 和其分布式存储 (SDS) 之间的数据路径中，但它们负责维护自身之间的关系，以便跟踪群集的状态。因此，MDM 至 MDM 流量对影响延迟的网络事件（例如 MLAG 中的物理网络链路丢失）很敏感。

MDM 是冗余的。因此，PowerFlex 不仅可以在延迟增加的情况下继续使用，而且在 MDM 丢失时也能继续使用。可以在托管 MDM 的节点上使用 MLAG。**但是，如果您需要在承载 MDM 至 MDM 流量的网络上使用 MLAG，请与 Dell EMC PowerFlex 代表合作，确保选择了采用双网络冗余的可靠设计，并将 MLAG 与原生 IP 级冗余相结合。**

在大多数情况下，建议 MDM 在两个或更多网络段（而不是 MLAG）上使用 IP 级冗余。MDM 可以共享一个或多个专用 MDM 群集网络。

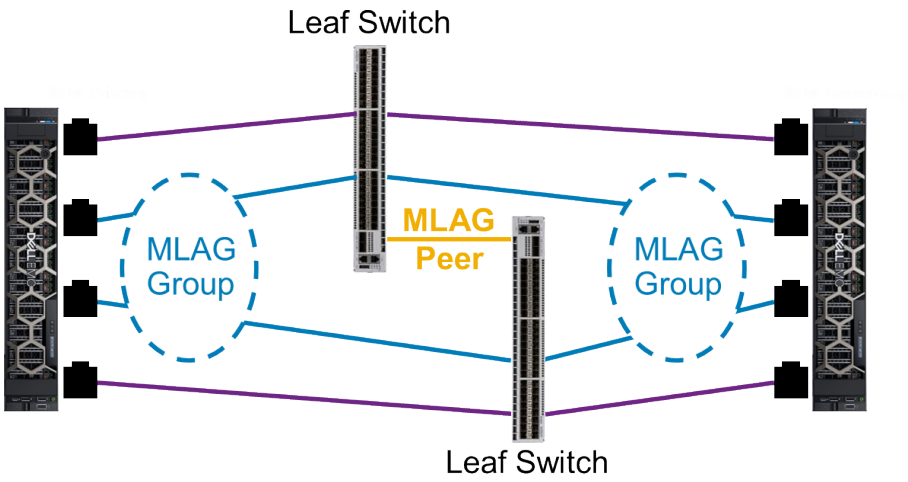

图 10　两个节点连接到两台分支交换机。MDM 流量应通过紫色链路，因为它们不在 MLAG 组中。

13 网络服务

13.1 DNS

MDM 群集维护系统组件及其 IP 地址的数据库。为了避免 DNS 中断影响 PowerFlex 部署，MDM 群集不通过主机名或完全限定域名 (FQDN) 来跟踪系统组件。如果在向 MDM 群集注册系统组件时使用主机名或 FQDN，则系统会将它解析为 IP 地址，并且使用其 IP 地址注册组件。

但在部署 VASA 提供程序和实施 vVol 时例外。在 PowerFlex 环境中使用 vVol 需要部署 PowerFlex VASA 提供程序（在单一模式或 3 节点群集中）。在 vSphere 环境中实施 vVol 技术需要 vCenter 服务器的完全限定 FQDN、使用 vVol 数据存储的 ESXi 主机和自托管的 VASA 提供程序。
所有这些组件必须使用有效的 DNS 解析。因此，使用的 DNS 服务必须高度可用，防止丢失 vVol 连接和功能。

总而言之，**除非实施了 vVol，否则主机名和 FQDN 更改通常不会影响 PowerFlex 部署中的组件间流量。**

14 WAN 上的复制网络

在使用 PowerFlex 原生异步复制时，还有其他因素需要考虑。在第 2.4 节和 3.9 节中，我们介绍了 Storage Data Replicator (SDR) 及其流量。在第 7.2.3 节中，我们介绍了额外的带宽要求。在本节中，我们考虑在广域网 (WAN) 上运行复制时的特定寻址和路由主题。给出的建议并不是具体的，因为实施详细信息取决于所使用的硬件和 WAN 拓扑。

14.1 附加 IP 地址

在保护域中，SDR 安装在与 SDS 相同的主机上，但 SDR 写入到日志卷的流量会发送到托管日志的所有 SDS，而不只是发送到在主机上与它同地协作的 SDS。在后端存储网络中，每个 SDR 侦听与 SDS 相同的节点 IP，因此应能够到达保护域中的所有 SDS。

但是，SDR 需要额外的不同 IP 地址，以便让它们能够与远程 SDR 通信。在大多数情况下，这些地址应该是具有正确配置网关的可路由地址。对于冗余，每个 SDR 应有两个。

14.2 防火墙注意事项

SDR 相互通信，并通过 TCP 端口 1088 在它们之间传送复制的数据。此端口必须对源系统端的任何防火墙中的出口开放，并且必须对目标系统端的入口开放。如果在两个系统之间的两个方向执行复制，则必须在防火墙中为两侧的出口和进口打开端口 1088。

14.3 静态路由

PowerFlex 异步复制通常在不共享相同地址段的物理远程群集之间的 WAN 上进行。如果默认路由本身不适合将数据包正确地发送到远程 SDR IP，则应配置静态路由，以便指示下一跳地址或出口接口或两者到达远程子网。

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

考虑一个每侧都有几个节点的小系统。每个节点有四个网络适配器，其中两个配置了用于 PowerFlex 群集内部通信的 IP，另两个配置了用于站点到站点外部通信的 IP 地址。

在本示例中，我们告诉节点通过指定的网关访问另一端的 WAN 子网。在源站点 A，网络接口 `enp130s0f0` 和 `enp130s0f1` 分别配置了 `30.30.214.0/24` 和 `32.32.214.0/24` 范围中的地址。我们可以为每个端口配置路由接口文件，以便通过指定的网关和接口为远程网络发送数据包。

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

用于远程网络 31.31.214.0/24 的数据包直接通过网关 IP 30.30.214.252 上的下一跳地址。对于发送到 33.33.214.0/24 的数据包也同样适用。

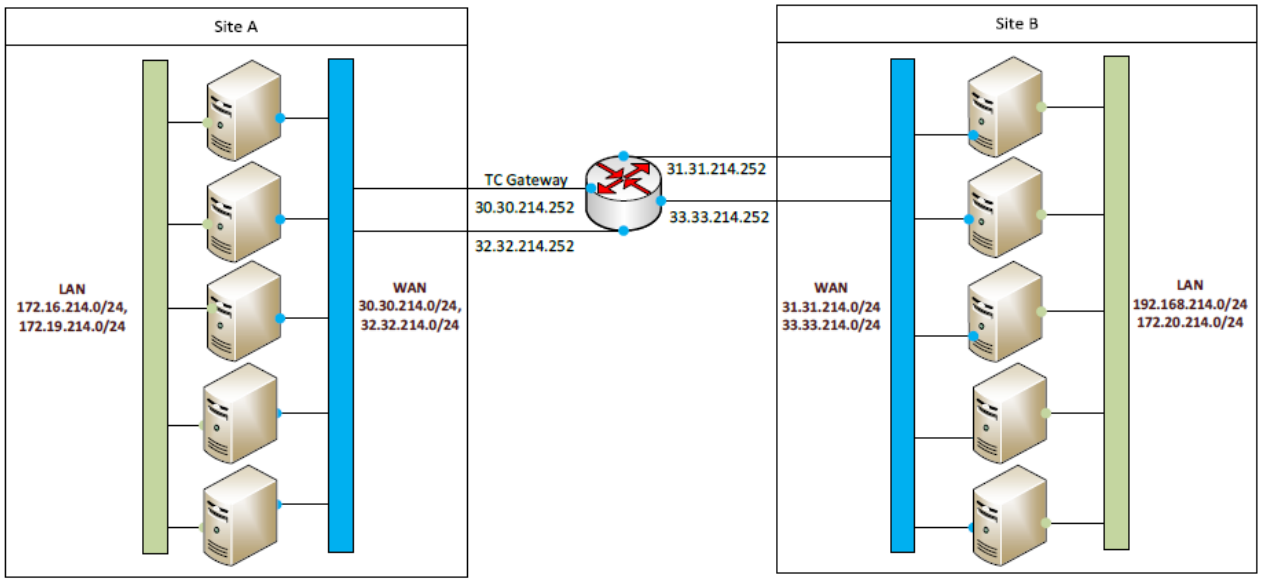

图 11　PowerFlex 复制的示例 WAN 拓扑。

静态路由配置的详细信息将因操作系统/虚拟机管理程序和整体网络体系结构而异，但一般原则是相同的。

14.4　MTU 和巨型帧

必须在 SDR 之间的网络接口上正确设置 MTU，以匹配 WAN 链路配置。在许多情况下，MTU 设置为 1500。如果在所有本地网络上都启用巨型帧以增强性能，那么记住这一点尤为重要。如果 MTU 与 WAN 配置不匹配时发生 IP 分段，则会导致复制性能降低。根据硬件配置，MTU 不匹配会导致数据包在到达接口时被完全丢弃。因此，在所有情况下，WAN 的 MTU 必须已知且经过测试。

15 动态路由注意事项

在包含数百个节点的大型分支-主干环境中，可能需要网络基础架构动态地路由 PowerFlex 流量。

路由 PowerFlex 流量的一个中心目标是减少路由协议的融合时间。当组件或链路出现故障时，路由器或交换机必须检测到故障；路由协议必须将更改传播到其他路由器；然后，每个路由器或交换机必须重新计算到每个目标节点的路由。如果正确配置了网络，这个过程可以在不到 300 毫秒的时间内完成：这个时间足够快，可以维持 MDM 群集的稳定性。

如果在极端拥塞或网络故障期间，融合时间超过 400 毫秒，则 MDM 群集可能会故障切换到辅助 MDM。如果 MDM 进行故障切换，系统将继续运行，且 I/O 将继续，但**为了使系统尽可能稳定，我们的目标是 300 毫秒。**其他系统组件通信机制的超时值要高得多，因此应针对要求苛刻的超时要求（MDM 的要求）来设计系统。

要获得尽可能最快的聚合时间，标准最佳实践适用。这意味着要遵循旨在实现这一目标的所有网络供应商最佳实践，包括不使用功率不足的路由器（弱链路），它们会阻碍快速融合。

在每个经过测试的网络供应商的默认 OSPF 或 BGP 配置中，融合时间均不够。**每个路由协议部署（无论是哪家网络供应商）必须包括性能调整，以便更大限度地缩短融合时间。**这些调整包括使用双向转发检测 (BFD) 和调整与故障相关的计时机制。

OSPF 和 BGP 均已使用 PowerFlex 进行测试。当路由协议和网络设备配置正确时，PowerFlex 可以在链路和设备故障期间正常运行，不会出现错误。但我们**建议使用 OSPF 而非 BGP。**当两者都经过优化配置以实现快速融合时，测试结果表明 OSPF 融合速度比 BGP 更快，这进一步证明这项建议可行。

15.1 双向转发检测 (BFD)

无论选择何种路由协议（OSPF 或 BGP），都需要使用双向转发检测 (BFD)。BFD 可降低与协议原生 hello 计时器相关的开销，从而快速检测链路故障。由于一些原因（包括路由器 CPU 和带宽利用率降低），BFD 提供比协议原生 hello 计时器更快速的故障检测。**因此，我们强烈建议使用 BFD，而不使用激进的协议 hello 计时器**

如果 PowerFlex 与 BFD 以及优化的 OSPF 和 BGP 路由一起部署，可在网络故障切换期间保持稳定。必须使用 BFD 实现亚秒级故障检测。

为了使网络融合，必须检测事件，将其传播到其他路由器，由路由器处理，并且必须更新路由信息库 (RIB) 或转发信息库 (FIB)。必须执行所有这些步骤才能实现路由协议融合，并且它们都应该在不到 300 毫秒内完成。

在使用 Cisco 9000 系列交换机的测试中，**BFD 保持 150 毫秒的计时器**就足够了。150 毫秒保持计时器的配置包含 50 毫秒传输间隔，具有 50 毫秒的 min_rx，乘数为 3。PowerFlex 建议使用 150 毫秒的最大保持计时器。如

果您的交换机供应商支持小于 150 毫秒的 BFD 保持计时器，则优先选择可实现的最短保持计时器。在可能的情况下，应以异步模式启用 BFD。

在使用 Cisco vPC (MLAG) 的环境中，**还应在所有路由接口和运行第一跳冗余协议 (FHRP) 的所有面向主机的接口上启用 BFD。**

```
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
```

图 12　Cisco 交换机上使用聚合-接入/主干-分支拓扑的 BFD 配置示例。BFD 配置为 150 毫秒的保持时间（间隔为 50 微秒；乘数为 3）。接口 port-channel51 上的 OSPF 和接口 Vlan30 上的 HSRP 均配置为 BFD 的客户端。

```
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
```

图 13　融合-接入拓扑中的戴尔 BFD 配置示例。BFD 配置为 150 毫秒的保持时间（间隔为 50 微秒；乘数为 3）。接口 port-channel51 上的 OSPF 和接口 Vlan30 上的 VRRP 均配置为 BFD 的客户端。

请注意关于这些配置的以下信息：

- 对于 port-channel 接口，必须启用每链路 BFD。
- 对于 BFD，必须禁用 IP 重定向。（使用覆盖来确保 BFD 正常工作）
- 只有接入/融合拓扑才需要 FHRP

15.2 物理链路配置

涉及链路故障的计时器是需要调整的候选项。链路断开和接口关闭事件检测和处理因网络供应商和产品线而异。**在 Cisco Nexus 交换机上，每个 SVI 接口上的“carrier-delay”计时器应设置为 100 毫秒，每个物理接口上的“link debounce”计时器应设置为 500 毫秒。**

载波延迟 (carrier-delay) 是交换机上的计时器。它适用于 SVI 接口。载波延迟表示在检测到链路故障时，交换机在通知应用程序之前应等待的时间。载波延迟用于防止在不稳定网络中出现摆动事件通知。在现代的分支-主干环境中，所有链路都应配置为点对点，从而提供稳定的网络。承载 PowerFlex 流量的 SVI 接口的建议值为 100 毫秒。

去抖动 (link debounce) 是一个延迟固件中的链路断开通知的计时器。它适用于物理接口。去抖动与载波延迟类似，但它适用于物理接口，而不是逻辑接口，并且仅用于链路断开通知。在等待期间停止流量。非零链路去抖动设置会影响路由协议的融合。对于承载 PowerFlex 流量的物理接口，链路去抖动计时器的建议值为 500 毫秒。

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

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

15.3 ECMP

需要使用等价多路径路由 (ECMP)。ECMP 在分支和主干交换机之间均匀地分发流量，并使用冗余的分支至主干网络链路提供高可用性。ECMP 与 MLAG 类似，但在第 3 层 (IP) 上运行，而不是在以太网上运行。

Cisco Nexus 交换机上的 OSPF 默认开启 ECMP。Cisco Nexus 交换机上的 BGP 默认不开启 ECMP，因此必须手动启用。所用的 ECMP 散列算法应为第 3 层 (IP) 或第 3 层和第 4 层（IP 和 TCP/UDP 端口）。

15.4 OSPF

OSPF 是优选路由协议，因为如果配置正确，它可以快速融合。使用 OSPF 时，分支和主干交换机全部驻留在一个 OSPF 区域。**为了提供稳定的 MDM 内通信，融合时间需要低于 300 毫秒。**在所有分支和主干交换机上，应将 OSPF 接口配置为点对点，并将 OSPF 进程配置为 BFD 的客户端。

这可以确保正确设置计时器；不改变默认值。**此外，对于 ToR-Agg（接入-聚合）拓扑中的第 3 层切换，OSPF 接口应配置为点对点。**

15.5 BGP

虽然 OSPF 可以更快融合，因而成为优先选择，但 BGP 也可以配置为在所需的时间段内融合。

默认情况下，Cisco Nexus 交换机上的 BGP 未配置为使用 ECMP。必须手动配置。默认情况下，IBGP 和 EBGP 都不支持 ECMP，并且必须进行配置。IBGP 的配置需要 BGP 路由反射器和添加路径功能，以便在主干和分支拓扑中完全支持 ECMP。

BGP 可以配置为每个分支和主干交换机代表不同的自动系统编号 (ASN)。在这种配置中，每个分支都必须与所有其他主干对等。

分支和主干交换机也应该允许交换机在多个 BGP 路径之间实现负载均衡，从而启用 ECMP。在 Cisco 交换机上，这包括将“`maximum-path`”参数设置为主干交换机的可用路径数。

采用 PowerFlex 的 BGP 需要在每个分支和主干邻近设备上配置 BFD。使用 BGP 时，**SDS 和 MDM 网络会通过分支交换机发出通告。**

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
```

图 14 Cisco Nexus 分支交换机（左）和主干交换机（右）上的 BGP 配置示例。它们驻留在相同的自动系统 (100)。调整“maximum-path”参数，以匹配要为 ECMP 使用的路径数。（在本示例中，路径数为 3，但情况可能并不总是这样）。为每个分支或主干邻近设备启用 BFD。分支交换机配置为向 PowerFlex MDM 和 SDS 网络发出通告

提醒：

- 在使用主干和分支拓扑结构的 PowerFlex 机架式系统上，BGP 用于控制平面的通信和 EVPN 的可访问性。OSPF 用于数据平面。
- Maximum-paths 允许多个 NVE 接口 VTEP 可访问性
- IBGP 配置为使用主干作为路由反射器
- 由于未使用 EBGP，BGP as-path multipath-relax 不适用

15.6 分支到主干带宽要求

假设存储介质不是性能瓶颈，则计算分支和主干交换机之间所需的带宽量涉及确定每台分支交换机至所连接主机的可用带宽量，如果 I/O 可能位于分支交换机本地，则减少带宽量，然后在每台主干交换机之间划分远程带宽要求。

考虑到在两个机架的情况下，每个机架都包含两台分支交换机和 20 台服务器，每台服务器有两个 25 Gb 接口，并且这些服务器中的每一台都双宿至机架中的两台分支交换机。在这种情况下，每台分支交换机的下行带宽计算方法为：

$$20\ \text{服务器} * 25\ \frac{\text{Gb}}{\text{服务器}} = 500\ \text{Gb}$$

每台分支交换机的下行带宽要求为 500 Gb。但是，某些流量在分支交换机对的本地，因此不需要通过主干交换机。

机架中的分支交换机的本地流量由配置中的机架数来确定。如果有两个机架，则可能有 50% 的流量在本地。如果有三个机架，则可能有 33% 的流量在本地。如果有四个机架，则可能有 25% 的流量在本地，依此类推。换句话说，可能是远程的 I/O 比例将是：

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

在本例中，有两个机架，因此有 50% 的带宽可能在远程：

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

在本例中，由于有两个机架，所以有 50% 的带宽可能在远程。将预期为远程的流量乘以每个分支交换机的下行带宽，得出每台分支交换机的总计远程带宽要求：

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

在本例中，使用 25 GbE 网络的分支交换机之间需要 250 Gb 带宽。但是，此带宽将在主干交换机之间分配，因此需要额外的计算。

因为远程负载在主干交换机之间平衡，所以如需了解从每台分支交换机到每台主干交换机的上行需求，请将远程带宽要求除以主干交换机的数量。

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

在本例中，通过融合主干交换机，每台分支交换机预计需要 250 Gb 的远程带宽。由于此负载将分布在主干交换机之间（假定有两台），因此，每台分支和主干交换机之间的总带宽计算方法为：

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

因此，对于无阻塞拓扑，每台分支和主干交换机之间两个 100 Gb 连接，总计 200 Gb 带宽就足够了。或者，可以在四个 40 Gb 连接中划分 125 Gb/s。

用于确定从每台分支交换机到每台主干交换机所需带宽量的公式可以概括为：

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

提醒：在实施复制的系统中，这些计算必须适应额外的后端复制存储流量。这可能会使这些示例中的需求增加一倍 — 分支交换机等需要四个 25 Gb 接口。

15.7 FHRP 引擎

对于在节点上使用 Cisco vPC 和 IP 级冗余的路由访问体系结构，戴尔建议为节点默认网关使用 FHRP。这样就可以在分支交换机发生故障时让默认网关故障切换至另一台分支交换机。FHRP 引擎因所用的交换机供应商而异。使用 Cisco 体系结构时，应使用 HSRP。对于戴尔交换机，应使用 VRRP。

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
```

图 15　一对 Cisco Nexus 聚合交换机上的 FHRP 引擎配置示例。活动的 vPC 对等应充当 FHRP 主设备，而备用 vPC 对等应充当 FHRP 辅助设备。

16 VMware 注意事项

虽然网络连接在 ESXi 中虚拟化，但本文档中所述的相同物理网络布局原则也适用。具体来说，这意味着除非已咨询 Dell EMC PowerFlex 代表，否则应避免在承载 MDM 流量的链路上使用 MLAG。

从运行 MDM 或 SDS 的虚拟机上的网络堆栈或 VMkernel 中 SDC 使用的网络堆栈的角度来考虑物理网络，这会很有帮助。考虑到来宾或主机级网络堆栈的需求，然后将其应用到物理网络，这样可以告知有关虚拟交换机布局的决策。

提醒：在版本 3.5 中，在基于 VMware 的超融合系统中尚不支持原生异步复制。因此，在本例中，上述基于 Linux 的系统的 IP 和吞吐量注意事项不会立即适用。但如果用户希望提前计划，则应考虑第 7.2.3 节中概述的额外吞吐量注意事项。

16.1 IP 级冗余

当使用双子网配置提供网络链路冗余时，需要两台独立的虚拟交换机。这是必需的，因为每台虚拟交换机都有自己的物理上行链路端口。当 PowerFlex 在超融合模式下运行时，此配置有 3 个接口：用于 SDC 的 VMkernel、用于 SDS 的虚拟机网络以及用于物理网络访问的上行链路。PowerFlex 原生支持此模式下的安装。

16.2 LAG 和 MLAG

使用 LAG 或 MLAG 时，需要使用分布式虚拟交换机。标准虚拟交换机不支持 LACP，因此不建议使用。使用 LAG 或 MLAG 时，在物理上行链路端口上进行绑定。

使用 vSphere 插件程序的 PowerFlex 安装原生不支持 LAG 或 MLAG 安装。而是可以在 PowerFlex 部署之前创建并在安装过程中选择。

如果运行 SDS 或 SDC 的节点已将链路聚合到交换机，则应将物理上行链路端口上的散列模式配置为使用“源和目标 IP 地址”或“源和目标 IP 地址以及 TCP/UDP 端口”。

如果需要，我们建议仅将其用作第二级冗余。

16.3 SDC

SDC 是用于实施 PowerFlex 存储客户端的 ESXi 的内核驱动程序。由于它在 ESXi 内核中运行，因此它使用一个或多个 VMkernel 端口与其他 PowerFlex 组件进行通信。我们重申实现本机 IP 级冗余的一般建议，在这种情况下，意味着每个 VMkernel 端口都映射到不同的物理端口。如果需要第二级冗余，则除了 IP 级冗余，还可以在分布式交换机层上实施 LAG 或 MLAG。

16.4 SDS

在 ESXi 上作为虚拟存储设备 (SVM) 的一部分部署 SDS。同样，我们建议的实施使用原生 IP 级冗余，每个子网分配给它自己的虚拟交换机和物理上行链路端口。如果需要第二级冗余，则除了 IP 级冗余，还可以在分布式交换机层上实施 LAG 或 MLAG。

16.5 MDM

在 ESXi 上作为虚拟存储设备 (SVM) 的一部分部署 MDM。强烈建议使用 IP 级冗余。**因此，单个 MDM 应使用两个或更多独立的虚拟交换机。**

17 虚拟化和软件定义的网络

在未来的更新中，我们有更多的内容要介绍。我们提供这些简短的说明，以明确对 SDN 支持的普遍误解。

17.1 Cisco ACI

我们不会对 Cisco ACI 上的 PowerFlex 提供直接或完全支持。具体来说，我们不支持 Cisco ACI 上的后端存储流量。但是，我们可以通过双网络扩展来支持它，在这种情况下前端客户流量流经 ACI 结构。

17.2 Cisco NX-OS

我们通过 NX-OS 独立软件支持 VxLAN EVPN 分支-主干结构。

18 验证方法

18.1 PowerFlex 原生工具

有两种主要的内置工具可以监控网络性能：

1. SDS 网络测试
2. SDS 网络延迟计量测试

18.1.1 SDS 网络测试

《Dell EMC PowerFlex v3.5 CLI Reference Guide》中介绍了 SDS 网络测试“start_sds_network_test”的使用。 要在运行后获取结果，请使用“query_sds_network_test_results”命令。

需要注意的是，应设置 parallel_messages 和 network_test_size_gb 选项，使其比运行测试的链路的最大网络带宽高出至少 2 倍。例如：一个 10 GbE NIC = 1250 MB * 2 = 2500 MB，或四舍五入为 3 Gb。在这种情况下，命令应使用参数“--network_test_size_gb 3”。这将确保在网络上发送足够的带宽以获得一致的测试结果。对于 25 GbE 网络配置，单个 25 GbE NIC = 3125 MB * 2 = 6250 MB，或 6 Gb。在这种情况下，命令应包括“--network_test_size_gb 6”。

并行消息大小应等于系统中的总核心数，最大配置为 16 个。

提醒：应在每个 SDS 上为每个配置的 SDS 网络运行此测试。

示例输出：

```
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
```

在上面的示例中，您可以看到从您正在测试的 SDS 到网络分段上的所有其他 SDS 的网络性能。确保每秒速度接近网络配置的预期性能。

18.1.2 SDS 网络延迟计量测试

“query_network_latency_meters”命令可用于显示 SDS 组件之间的平均网络延迟。SDS 组件之间的低延迟对于出色的写入性能至关重要。如果使用的是 10 千兆网络或速度更快的网络连接，则在运行此测试时，查找高于几百微秒的异常值和延迟。

提醒：此测试应从每个 SDS *并*在每个 SDS 网络上运行。

示例输出：

```
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 和 Tracepath

提醒：在配置 PowerFlex 之前，应使用 Iperf 和 NetPerf 来验证您的网络。如果发现 Iperf 或 NetPerf 有问题，则可能存在需要调查的网络问题。如果未发现 Iperf/NetPerf 问题，请使用 PowerFlex 内部验证工具进行额外和更准确的验证。

Iperf 是一个流量生成工具，可用于测量 IP 网络上最大可能带宽。使用 Iperf 功能集可调整各种参数以及有关带宽、丢失和其他测量值的报告。使用 Iperf 时，应使用多个并行客户端线程运行。每个 IP 插槽八个线程是一个不错的选择。

NetPerf 是一个基准，可用于衡量许多不同类型网络的性能。它提供用于单向吞吐量和端到端延迟的测试。

可以使用 Linux “`tracepath`” 命令来发现路径上的 MTU 大小。

18.3 网络监控

必须监控网络的运行状况，以确定任何会阻碍您的网络以更适合的容量运行的问题，并防止网络性能下降。市面上有大量可供使用的网络监控工具，可提供许多不同的功能集。

Dell Technologies 建议监控以下方面：

- 输入和输出流量
- 错误、丢弃和溢出
- 物理端口状态

18.4 网络故障处理基础知识

- 使用 ping 验证 SDS 和 SDC 之间的端到端连接
- 在两个方向上测试组件之间的连接性
- SDS 和 MDM 通信不应超过 1 毫秒的仅网络往返时间。
- 使用 ping 验证组件之间的往返延迟
- 检查交换机侧的端口错误、丢弃和溢出
- 验证 PowerFlex 节点是否启动
- 验证是否在所有节点上安装并运行 PowerFlex 流程
- 检查所有交换机和服务器上的 MTU，尤其是在使用巨型帧时
- 验证用于站点到站点 SDR 通信的 MTU 是否适合 WAN
- 验证站点到站点 SDR 通信的静态路由配置，并测试 WAN 上的端到端连接
- 如果可能，可选择 25 千兆以太网或更高速度的以太网代替 10 千兆以太网
- 检查操作系统事件日志中的 NIC 错误、高 NIC 溢出率 (> 2%) 和丢弃的数据包
- 检查 IP 地址，确认没有有效的 NIC 关联
- 确认网络或节点不会阻止 PowerFlex 所需的网络端口
- 使用事件日志或操作系统网络命令检查运行 PowerFlex 的操作系统上是否有数据包丢失情况
- 确认节点上运行的任何其他应用程序都未尝试使用 PowerFlex 所需的 TCP 端口
- 将所有 NIC 设置为全双工，开启自动协商，设置您的网络支持的最大速度

- 检查 PowerFlex 原生工具测试输出
- 检查 RAID 控制器配置错误（这与网络无关，但这是常见的性能问题）
- 如果遇到问题，请在被覆盖之前尽快收集日志
- 《Troubleshoot and Maintain Dell EMC PowerFlex v3.5》和《PowerFlex v3.5 Log Collection Technical Notes》中提供了额外的故障处理、日志收集信息和常见问题解答。

19 总结

所选的部署选项、网络拓扑、性能要求、以太网、动态 IP 路由和验证方法，所有因素都融入到强大且可持续的网络设计中。Dell EMC PowerFlex 群集最多可扩展到包含各种节点类型、存储介质和部署配置的 1024 个节点，因此应该根据未来的增长来调整网络安装的规模。PowerFlex 可以部署在超融合模式下，其中计算和存储驻留在同一组节点，也可以部署在双层模式下，存储和计算资源分开，这个事实也会影响您的决策。为了实现卓越的性能、可扩展性和灵活性，网络的设计必须考虑到业务的需求。遵循本指南中的原则和建议，打造具有弹性、可大规模扩展和高性能的块存储基础架构。