DELL Technologies

技术白皮书

Dell EMC PowerFlex：复制简介

PowerFlex 复制概述和基本配置

摘要

Dell EMC PowerFlex™ 软件定义的基础架构提供原生异步复制。本白皮书概述了 PowerFlex 复制技术、部署和配置的详细信息以及复制 PowerFlex 群集的设计注意事项。

2021 年 6 月

修订记录

日期	描述
2020 年 6 月	初版
2021 年 5 月	更新了日志容量和网络建议
2021 年 6 月	适用于 PowerFlex 版本 3.6 的更新

致谢

内容所有者：Brian Dean，PowerFlex 技术营销

支持：Roy Laverty、Matt Hobbs、Neil Gerren

目录

执行摘要

PowerFlex™ 软件定义的基础架构平台提供非凡的灵活性、弹性和简易性，同时大规模实现可预测的性能和抗风险能力。随着 PowerFlex 不断发展，增加了原生异步复制功能，使随附的企业级存储服务集得以扩展。客户需要灾难恢复和复制功能来满足业务和法规遵从性要求。复制也可以用于其他应用场景，例如，分流要求苛刻的分析工作负载，将其与其他业务关键型系统的任务关键型工作负载隔离开来。本白皮书涵盖：

- PowerFlex 复制的核心设计原则
- 配对存储群集的配置要求
- 复制一致性组的配置要求
- 网络注意事项
- 复制应用场景

1 简介

PowerFlex 是一个软件定义的存储平台，旨在降低运营和基础架构复杂性，通过提供灵活性、弹性和简易性并大规模地提供可预测的性能和抗风险能力，使组织能够更快地发展。PowerFlex 系列软件定义的基础架构为在托管统一构造中结合计算与高性能存储资源奠定了基础。由于具有多个平台部署选项（例如，机架、设备或就绪型节点），所有这些都提供服务器 SAN、HCI 和仅存储体系结构，因此带来了灵活性。

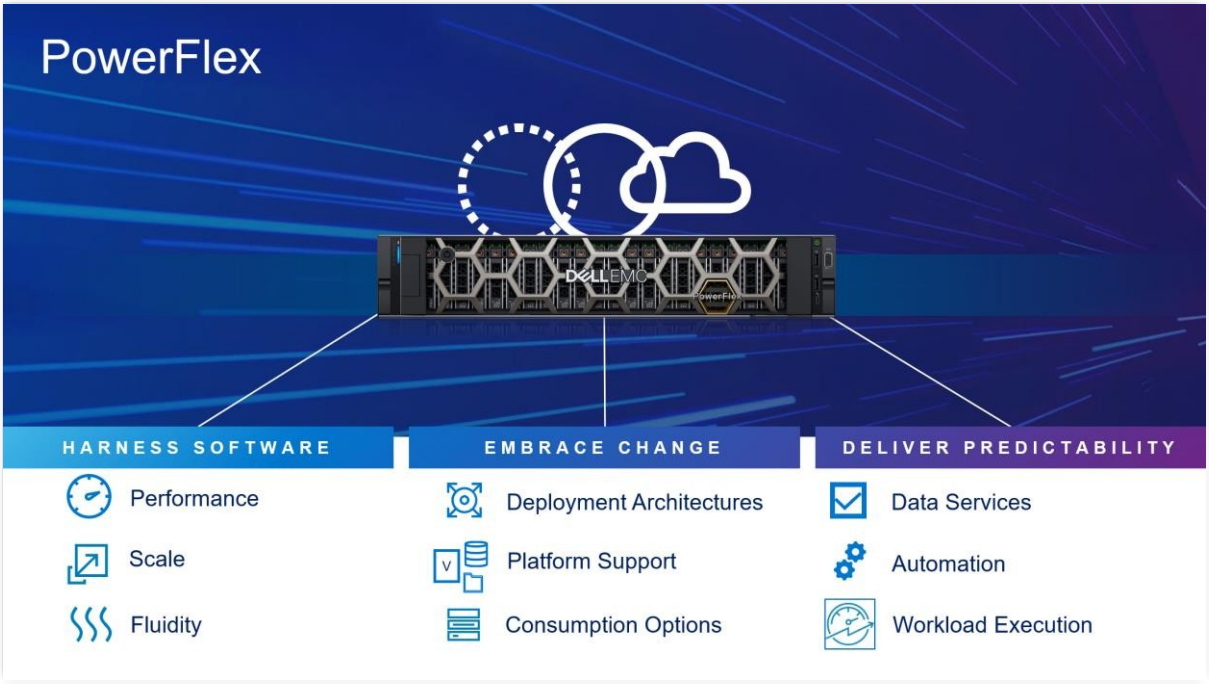


图 1 PowerFlex 概述

PowerFlex 提供了一系列应用程序部署所需的灵活性和可扩展性，无论它们是在裸机上、虚拟化还是容器化。

它提供了要求苛刻的企业所需的性能和弹性，展示了 99.9999% 或更高的任务关键型可用性，以及稳定和可预测的延迟。

PowerFlex 轻松提供数百万 IOPs，延迟达到亚毫秒级，非常适合高性能应用程序，并且非常适合需要一个可与公有云和混合云协同作用的灵活基础的私有云。对于将异构资产整合到一个具有灵活、可扩展体系结构（利用该体系结构可以自动管理存储和计算基础架构）的单一系统中的组织，它也非常适合。

2 PowerFlex 异步复制体系结构

要了解复制的工作原理，我们必须首先考虑 PowerFlex 本身的基本体系结构。

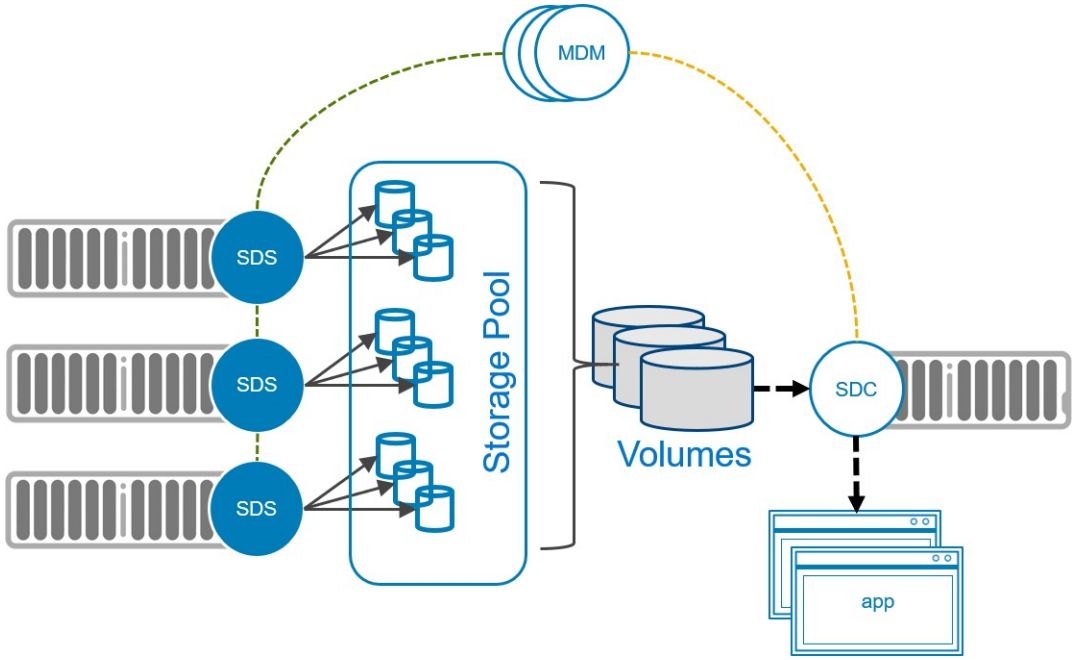


图 2 PowerFlex 基本组件和体系结构

向存储群集提供介质的服务器运行 Storage Data Server (SDS) 软件元素，允许 PowerFlex 聚合介质，同时将这些资源作为创建逻辑卷的一个或多个统一池来分享。

使用存储的服务器运行 Storage Data Client (SDC)，从而可以通过主机 SCSI 层提供对逻辑卷的访问。请注意，不使用 iSCSI，而是使用 TCP/IP 存储网络上运行的弹性负载管理、负载均衡网络服务。

Metadata Manager (MDM) 控制通过系统但不在数据路径中的数据流。相反，它会在 SDS 群集中创建和维护有关卷分布的信息，并将映射分发到 SDC，告诉它在何处放置和检索地址空间每个部分的数据。

这三个基本元素构成现今出色的软件定义存储解决方案的基本部分，可以线性扩展到数百个 SDS 节点。

在考虑用于复制的体系结构选项时，保持 PowerFlex 的可扩展性和弹性，这很重要。PowerFlex 中的复制体系结构是刚才说明的基本部分的自然扩展。

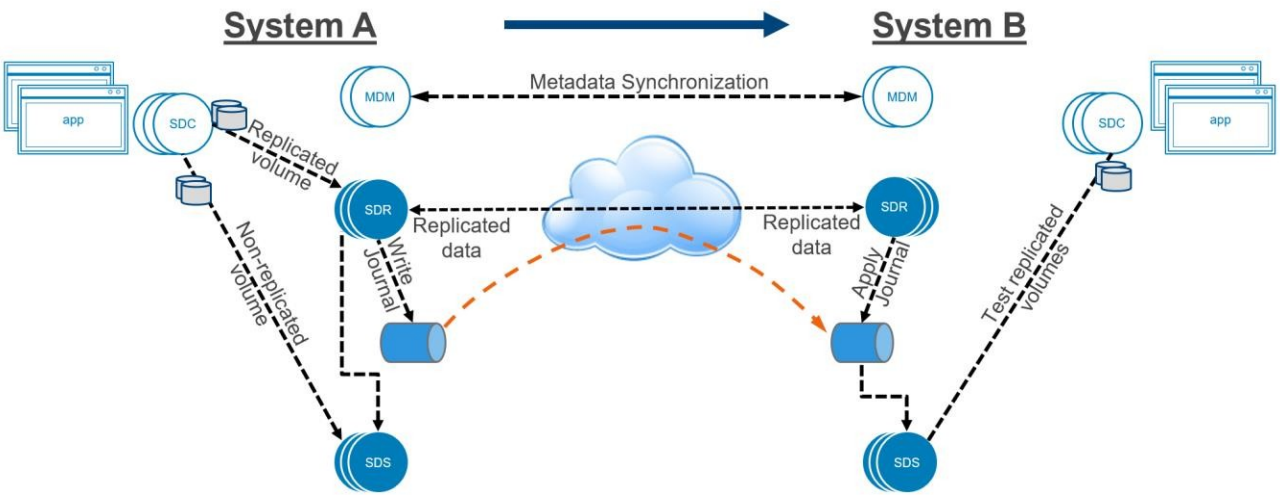


图 3　PowerFlex 简化的复制体系结构

PowerFlex 版本 3.5 引入了一个称为 Storage Data Replicator (SDR) 的新存储软件组件。图 3 描绘了 SDR 在整体 PowerFlex 复制体系结构中的位置。它的作用是代理 SDC 和最终存储数据的 SDS 之间的复制卷的 I/O。写入 I/O 拆分开来，将一个副本发送到目标 SDS，另一个副本发送到复制日志卷。
从 SDS 的角度来看，SDR 位于 SDS 和 SDC 之间，看起来好像是 SDC 正在发送写入。（但从网络产品的角度来看，SDR → SDS 流量仍为后端/存储流量。）与之相反，对于 SDC 来说，SDR 看起来好像是一个 SDS，可以向其发送写入。

SDR 仅传递复制卷的流量。（实际上只传递主动复制卷的流量；下面会介绍它们的细微差别）。如往常一样，非复制卷 I/O 在 SDC 和 SDS 之间直接流动。与往常一样，MDM 会指示每个 SDC 在何处读取和写入数据。由 MDM 向 SDC 提供的卷地址空间映射确定卷的数据发送到何处。但 SDC 不知道写入目的地是 SDS 还是 SDR。SDC 不知道复制。

2.1　写日志与快照创建

关于如何实施复制，有两种观点。许多存储解决方案采用基于快照的方法。使用快照，很容易识别两个时间点之间的块更改增量。但是，随着恢复点目标变得更小，所需的快照数量会大幅增加，这会严格限制 RPO 可以有多小。相反，PowerFlex 使用基于日志记录的方法。

基于日志的复制可实现非常小的 RPO，重要的是，它不受系统中或给定卷上可用快照的最大数量的限制。

检查点（或间隔）保留在日志中，并且这些日志以 PowerFlex 卷的形式存在于同一个保护域的存储池中。但是，日志卷不需要驻留在与待复制的卷相同的存储池中。日志卷会随着写入的提交、传送、确认和删除而动态地调整大小。因此，日志缓冲区使用的实际容量会随时间发生变化。

2.2 日志容量预留

虽然实际容量因使用情况而异，但必须手动设置日志预留（指定允许复制过程使用的最大容量）。适当调整日志卷预留空间对 PowerFlex 群集的运行状况至关重要，尤其是在 WAN 中断和其他故障情况下。例如，即使 SDR 不能将日志间隔传送到目标站点，日志卷也必须有足够的可用容量来继续接收复制数据。（当然，对于不使用复制的 PowerFlex 的安装，不需要任何日志空间预留。）如果无法传送日志间隔，日志缓冲容量将增加，可能会完全填满。因此，您必须考虑中断时可能会发生的最大累积写入。如果日志缓冲区空间完全填满，则复制对卷将需要重新初始化。下面介绍了关于此主题的更多内容。

管理员设置并调整日志卷的最大预留大小。日志容量的最低要求是 28 GB 乘以 SDR 会话数，其中 SDR 会话数等于已安装的 SDR 数加一。但是，需要进行一些额外的计算，因为预留大小在系统中表示为包含每个日志卷的存储池的*百分比*。一般来说，为复制日志预留存储池容量的至少 5%。

日志的预留容量可能会在多个存储池中拆分为多个卷，或者复制日志可能全部驻留在保护域的一个存储池中。**日志卷所驻留的任何存储池的性能必须匹配或超过复制卷所驻留的任何存储池的性能要求。**

虽然日志容量必须足以适应各种因素，例如卷开销、空闲空间预留（以便承受节点故障或适应受保护的维护模式）等，但更重要的一点是可能的 WAN 中断。如果我们考虑到这种情况，我们最终会考虑所有其他考虑因素。

首先，评估每个应用程序所需的日志容量。由于应用程序 I/O 会随着时间的推移而发生变化，因此我们需要知道应用程序在最繁忙时段的最大写入带宽。最小中断限额为 1 小时，但我们强烈建议在计算中使用三小时。

计算示例

- 我们的应用程序在高峰时段生成 1 GB/s 的写入
- 使用 3 小时作为可承受的中断，我们用 10800 秒来计算
- 所需的日志容量预留为 1 GB/s * 10800 s = ~10.547 TB
- 由于日志容量按存储池容量的百分比来计算，因此我们将所需的空间除以存储池可用容量。假设该容量是 200 TB。
- 100 * 10.547 TB/200 TB = 5.27%。
  - 为了留出安全裕度，我们将把它四舍五入为 6%。

对要复制的每个应用程序重复此操作。

提醒：随着存储池的大小和容量发生变化，该百分比也会改变。由于池容量发生变化，所以需要重新调整日志预留。管理员可以随时通过 UI、CLI 或 API 调整日志容量预留百分比。

2.3 日志间隔和数据流

每个群集都可以是复制源和目标。这使客户能够在区域分开的群集之间拆分应用程序，同时确保应用程序在任一位置可用。

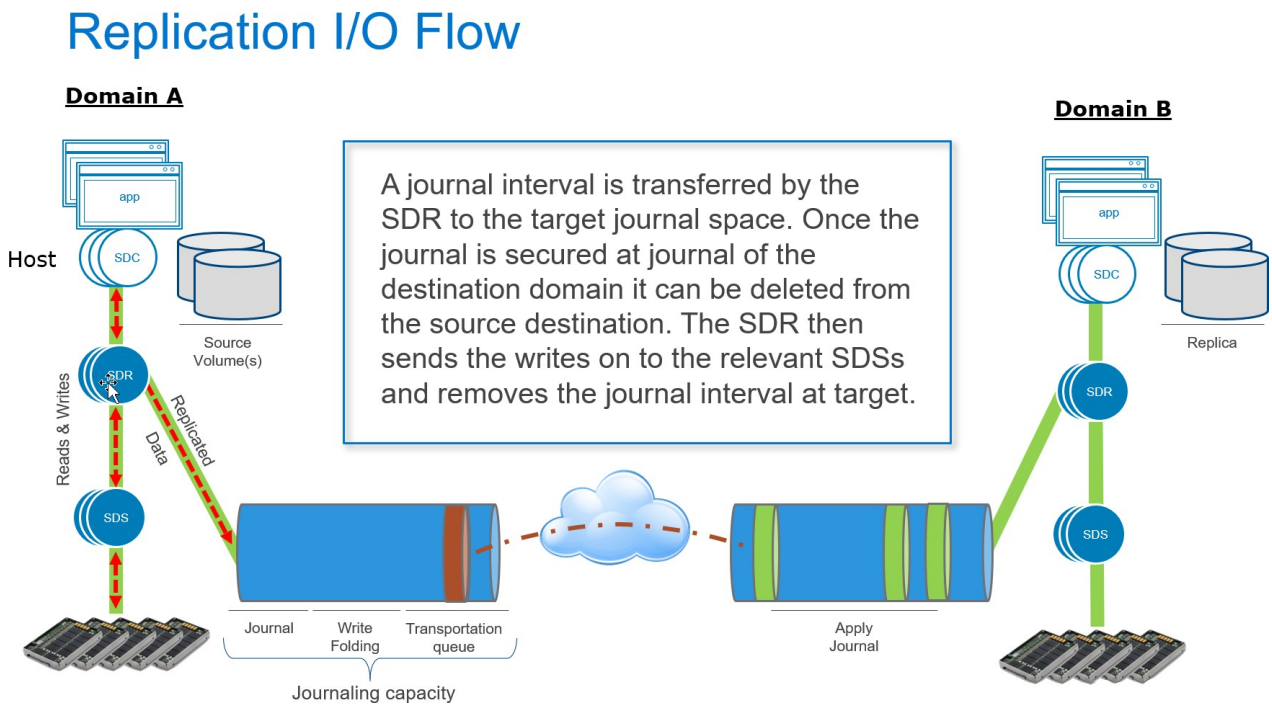


图 4　PowerFlex 简化的复制 I/O 流

源 SDC 的卷映射向 SDR 发送复制数据的写入，这会复制写入并转发。本地 SDS 正常处理这些写入，而 SDR 会汇集包含检查点的日志文件以保持写入顺序。

日志在源系统上的日志缓冲区中进行批处理。当它们靠近队列的头部时，系统会对它们进行扫描，并且合并（合拢）重复块写入，以便更大限度地减少通过网络发送的数据量。

由 SDR 通过本地网络或分配给复制的外部 WAN 网络上的专用子网将日志间隔发送到远程目标日志缓冲区，一旦在目标日志中得到确认，就会从源中删除。

在目标系统上，日志由 SDR 处理并应用到目标卷，从而将写入内容传递到相关 SDS。SDS 像往常一样管理主要副本和辅助副本。

这就提出了一个常见问题：*复制卷中的压缩受到怎样的影响？* 简短的回答是压缩数据不通过 WAN 发送。SDR 是 SDC 与 SDS 之间的传递者，它在压缩方面不起任何作用。压缩由 SDS 完成，SDS 接收写入并将其存储到运行 SDS 的主机的本地磁盘。对于给定 SDS，PowerFlex 中的压缩在本地进行。因此，系统不会压缩通过网络发送的数据。

目标端的 SDR 接收来自目标端 SDS 的确认后，它会继续进行待处理日志间隔中包含的下一个写入。在处理并确认日志间隔中的上一个写入后，删除间隔，并且日志容量可供重复使用。

还有几个其他 SDR 子流程协同工作以保护数据的完整性，但这个说明涵盖了所有基本部分。

需要注意与卷迁移相关的一个限制。无法将复制卷从一个保护域迁移到另一个保护域。这是因为复制日志不会跨保护域。

3 部署和配置 PowerFlex 群集以进行复制

在部署任何新的 PowerFlex 群集之前，必须执行适当的系统和存储规模调整。复制增加了额外的规模调整问题。Dell Technologies 技术销售人员可以访问系统规模调整实用程序，它接受包括工作负载特性、复制占用空间、WAN 带宽和质量以及网络设计和基础架构在内的输入。

在添加异步复制时，还需要考虑额外的群集设置要求。我们需要

- 为参与复制的群集提供一种安全通信的方法。
- 将卷对分组为一致性组。
- 测试失败的方法，或甚至在不影响主应用程序的情况下分发工作负载。
- 配置物理 WAN 网络，以便在目标群集位于另一个数据中心时进行外部复制
- 复制活动的额外 IP 地址

我们将在本章中介绍所有这些主题。

3.1 部署和配置

当部署要与复制一起使用的群集对时，需要执行一些必需的配置步骤。

3.1.1 存储群集证书颁发机构根证书的交换

PowerFlex 系统根 CA 证书必须在复制群集之间交换，以防止可能的安全攻击。由于这是一个与安全相关的问题，使用 PowerFlex 命令行界面来执行此步骤。在每个系统上，创建一个证书并将其发送到复制对中的另一台主机。示例命令：

```
scli --extract_root_ca --certificate_file /tmp/sys0.cert
```

从群集提取证书。然后将该证书手动复制到合作伙伴群集。为了导入证书，我们在合作伙伴系统上使用以下形式的命令：

```
scli --add_trusted_ca --certificate_file /tmp/sys0.cert --comment Site-A
```

在两个系统上生成、交换和导入证书后，证书交换步骤就完成了。

3.1.2 对等存储群集

配置复制卷对之前需要执行的下一步是对等。对等建立了系统之间的数据路径和通信。这可以使用 PowerFlex WebUI 来完成，但我们首先需要一条关键信息。我们将使用 PowerFlex CLI 来捕获两个存储群集的系统 ID。只需登录到 PowerFlex CLI 即可找到它们。通过 CLI 对群集进行身份验证的操作可显示群集 ID。您将需要源和目标系统的 ID。

```
[root@tme-102T-9 ~]# scli --login --username admin
Enter password:
Logged in. User role is SuperUser. System ID is 7dda10f5693d3f0f
```

图 5　捕获 PowerFlex 系统 ID

要开始对等，请浏览至 PowerFlex WebUI 中的 PROTECTION 侧边菜单，然后单击它。（此示例使用版本 3.6 UI，与 3.5 UI 略有不同，但在 3.5.x 中仍然可以轻松地执行该示例。）

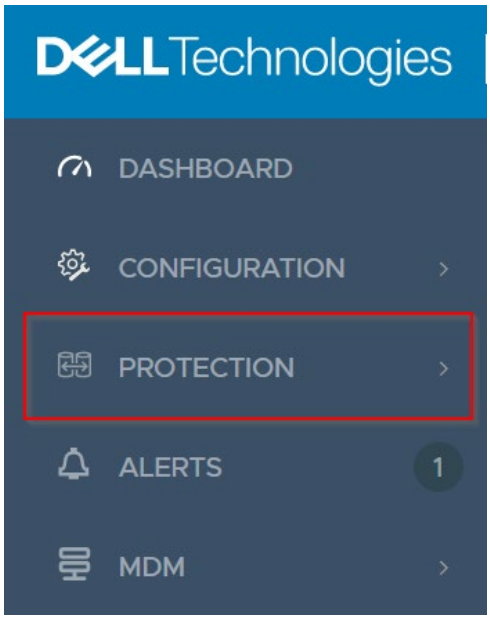


图 6

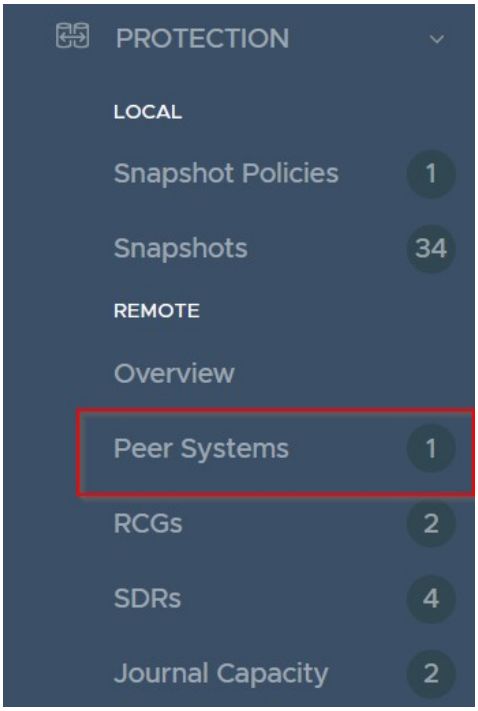


图 7

在 Replication 菜单项下，选择 **Peer Systems**。

要添加系统对等项，请单击 **ADD** 按钮。

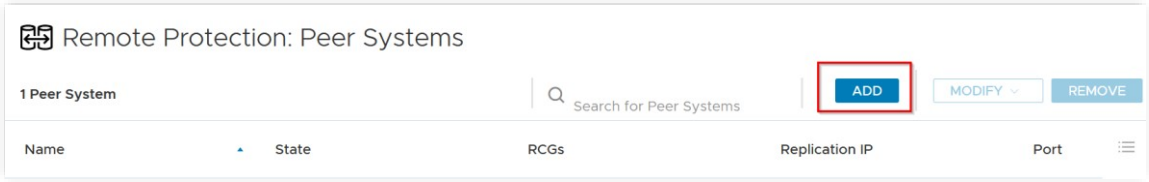


图 8　添加对等

填写目标群集主 MDM 的名称、远程系统 ID 和 IP。单击 **Add IP**。如有必要，添加额外 IP。单击 **Add Peer** 完成向导。

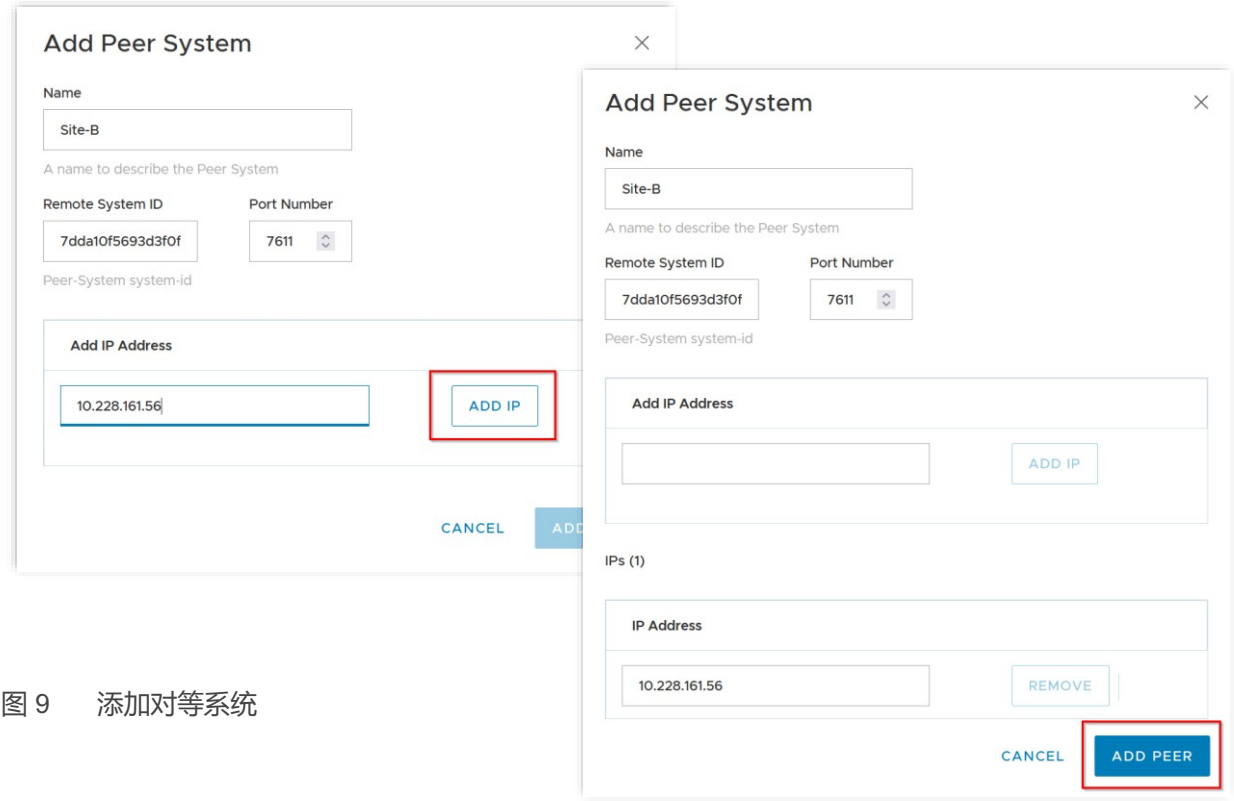


图 9　添加对等系统

添加对等后，使用相同的步骤在目标存储群集上重复该过程，然后输入主群集的远程系统 ID。完成后，系统在两个方向上实现对等，并且您已经准备好开始对复制的卷进行配对。

3.2　复制一致性组

复制一致性组 (RCG) 确立一个或多个卷对的复制的属性和行为。其中一个此类属性是目标复制存储群集。虽然给定的 RCG 只能复制到一个目标群集，但原则上，其他 RCG 可以复制到其他群集，前提是它们已交换证书并已对等。但是，在 PowerFlex 原生异步复制的 3.5 和 3.6 版本中，一个源站点只能与一个其他站点对等。未来版本将允许额外的复制拓扑。

在创建 RCG 之前，源和目标系统上都必须有我们的复制卷对，并且其大小必须相同。在 PowerFlex 版本 3.5.x 和 3.6 中，必须手动创建目标卷。虽然卷大小必须完全相同，但不需要驻留在同一类型的存储池中（MG 与FG 存储池），它们也不需要有相同的属性（非精简与精简、压缩与未压缩）。如果必须调整卷的大小，应先扩展目标卷。以这种方式扩展卷可防止复制出现中断。这意味着必须知道要复制哪些卷，以便在数据超出源卷时可以遵循此做法。

RCG 非常灵活。对于某些应用场景，您可以将与应用程序关联的所有卷分配到单个 RCG。对于较大的应用程序，您可以根据数据保留、数据类型或相关应用程序静默程序创建多个 RCG，以便在需要时启用读取一致性快照。一般而言，RCG 为崩溃一致性。如果在创建快照时遵循应用程序静默规则，则快照可以设置为读取一致性。这对存储平台没有特殊要求，但通常需要对应用程序编写脚本。

在 RCG 配置中指定**恢复点目标**。如图 10 中所示，PowerFlex 版本 3.6 RPO 可以设置为 15 秒至 60 分钟。（提醒：在 PowerFlex 版本 3.5.x 中，可用的最小 RPO 为 30 秒。）

要创建 RCG，请登录到 WebUI 并导航至 **PROTECTION > REMOTE > RCGs**，然后单击 **ADD** 按钮。

这是创建 RCG 的第一步，在这一步要提供以下信息：

- RCG 的名称
- 所需的 RPO
- 源保护域
- 目标系统
- 目标保护域

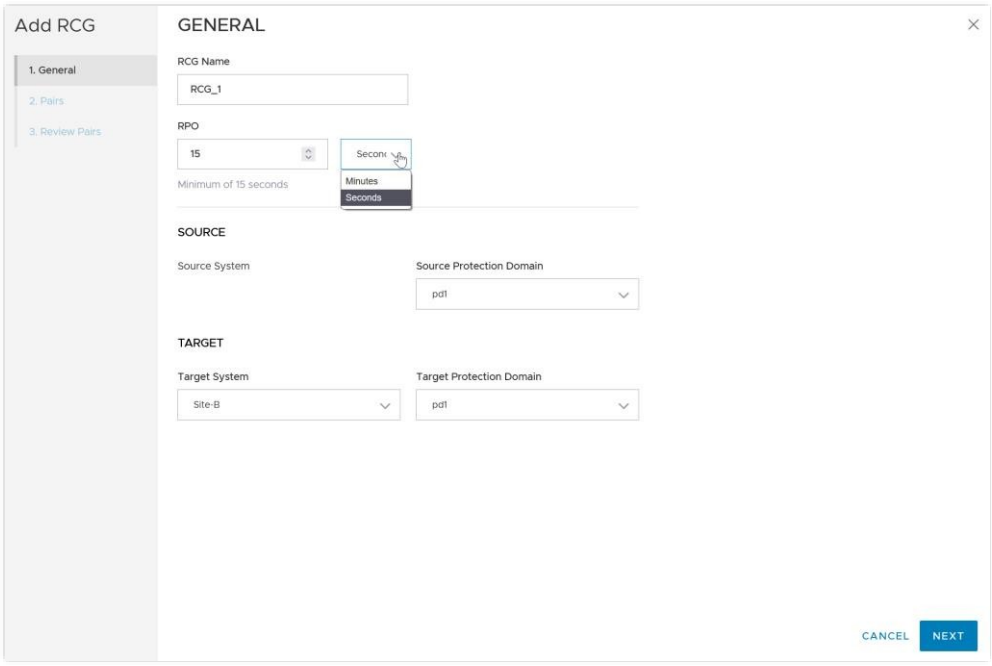


图 10　添加 RCG — 设置 RPO

为了完成此操作，我们首先匹配所需的源和目标卷。

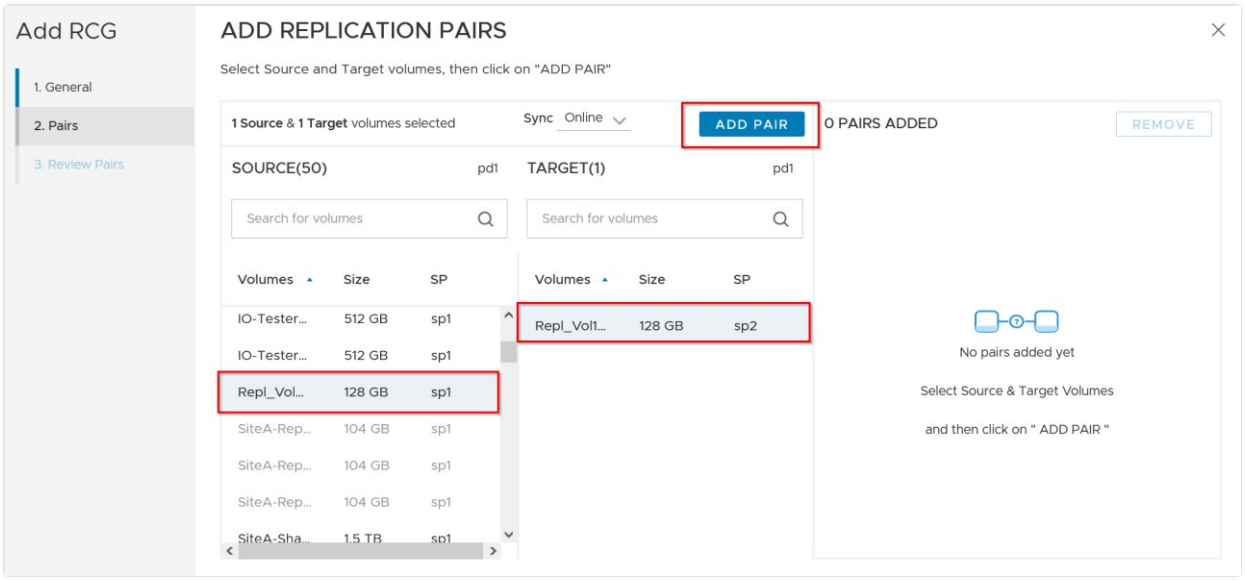
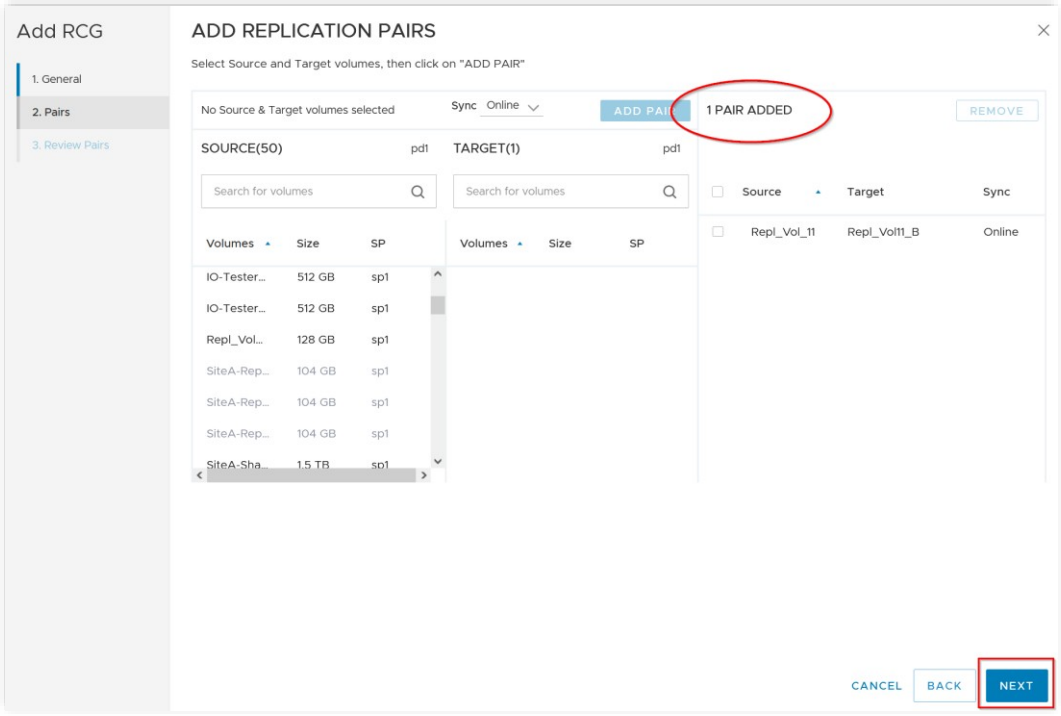


图 11　添加 RCG — 匹配源和目标卷

- 在源和目标列表中，未配对的卷以暗色文字显示。选择源卷时，系统会显示未配对的目标卷，前提是它们具有与所选卷相同的容量。选择两个卷后，单击 **ADD PAIR** 按钮，将卷对移动到右侧出现的列表中。添加完所有卷对后，单击 **NEXT** 按钮继续。



- 然后会显示一个摘要，您可以在其中选择卷对，也可以在需要时删除它们。

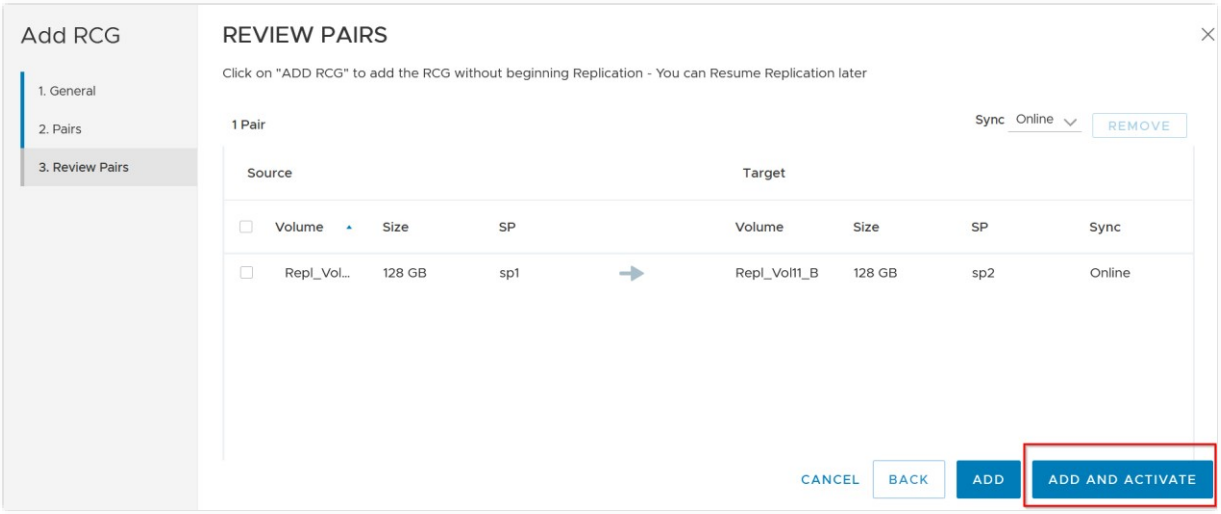


图 12　添加 RCG — 查看配对

- 最后一次单击鼠标后，我们可以选择创建 RCG 并立即激活卷同步，也可以直接添加配置，而不激活。我们将在下面更详细地讨论活跃和不活跃 RCG 状态。

- 您可以随时在 RCG 中添加卷或从中删除卷。出于对初始同步期间过多 I/O 的担心，并且根据卷的大小，您可以选择在 RCG 首次创建时一次仅添加单个卷对，但这通常是不必要的。

4 复制监控和配置

4.1 复制控制面板

PROTECTION → Remote → Overview 区域提供了一个控制面板，用于确定系统中复制的整体运行状况和状态。

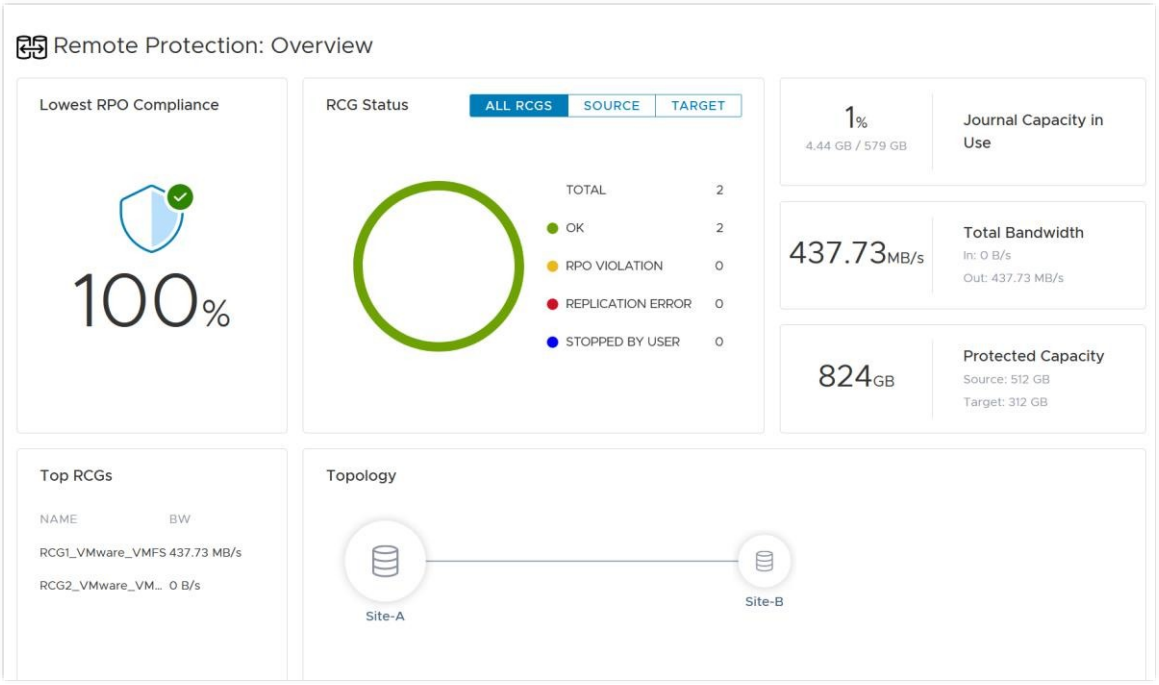


图 13 复制概述

4.2 复制一致性组视图

- 使用 WebUI 中的 **PROTECTION → Remote → RCGs** 视图来监测各个复制一致性组的运行状况和状态，或添加新的一致性组。

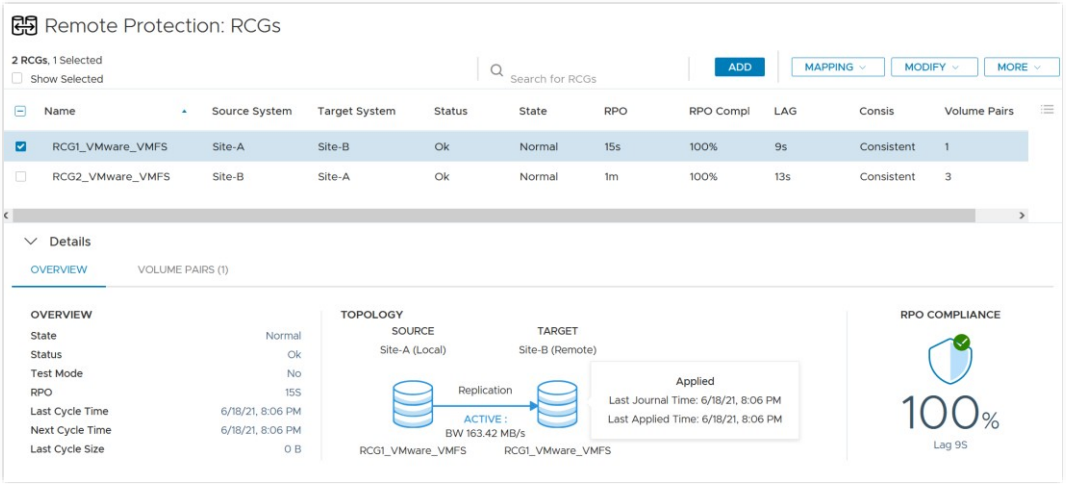


图 14 RCG 概述

选择任意 RCG 的行将打开详细信息窗格，在这里可以查看 RCG 及其组件卷的状态。

单击 RCG 的复选框以选择它，即可启用操作菜单。在 MORE 菜单下，我们可看到以下选项。

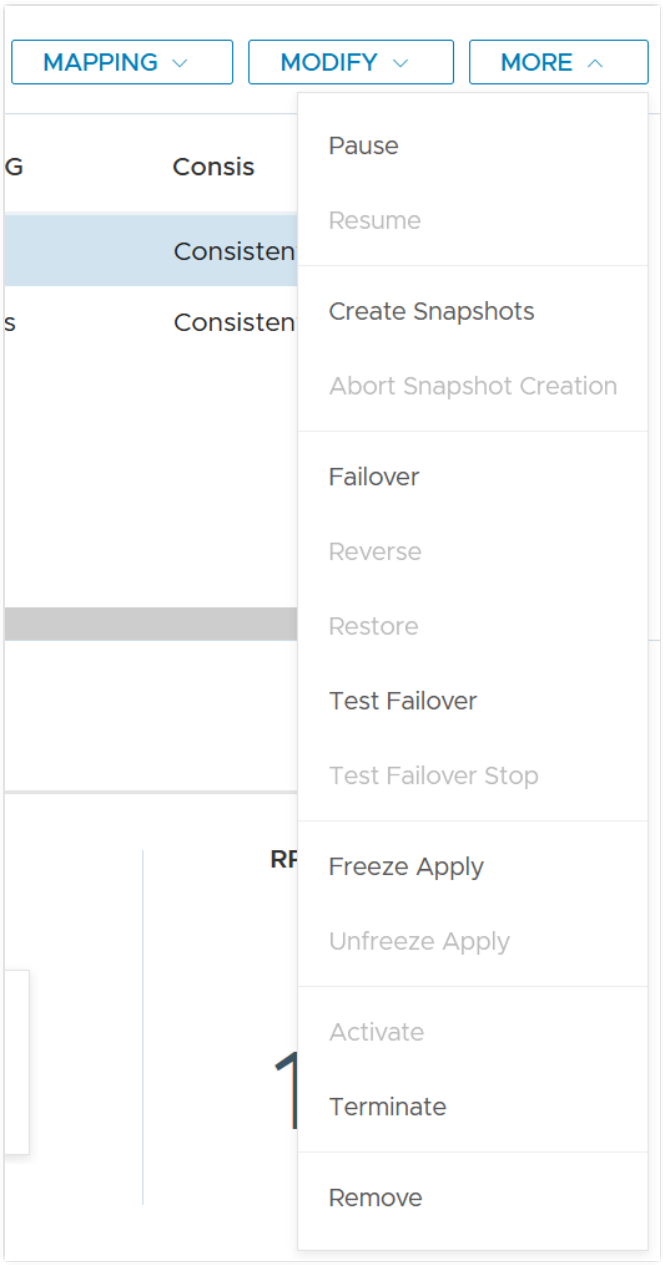


图 15　RCG 管理选项

Pause：此操作会暂停源和目标之间的复制。这可以防止在恢复复制之前将日志发送到目标群集。在源日志卷中仍会收集对复制卷的写入。

Create Snapshots：在目标系统的 RCG 中生成每个卷的快照。这对于远程测试应用程序或 DR 活动非常有用。没有用于管理或删除快照的 RCG 菜单选项，因此必须在目标系统的快照列表中对其进行映射/取消映射和手动删除。

Failover：强制执行故障切换事件，将 RCG 内的卷的主要所有权传递到目标系统。这也会将源端卷上的主机访问配置文件切换为只读，以及将目标上的访问配置文件切换为读/写。完成此操作后，对于计划的故障切换，您还可以选择 **Reverse** 命令来恢复 RCG 卷的保护，只是现在方向相反。如果您希望中止故障切换操作，请选择 **Restore** 选项，恢复为原始复制状态和方向。

Test Failover：这会在目标系统上自动创建快照，并将原始目标卷映射替换为快照的映射。您可以使用此命令对卷执行写入测试，同时防止测试活动损坏源卷。

Freeze Apply：此选项会冻结应用程序，禁止写入到目标卷的目标日志中。这不会暂停站点间的复制，并且目标系统的应用日志卷中会累积日志间隔。完成后，选择 **Unfreeze Apply** 将应用程序恢复到目标卷。

Activate/Terminate：这些是 PowerFlex 版本 3.6 中的新选项。如果创建了 RCG 但未激活，或已置于非活动状态，则可以在此处激活它。激活将启动所有与复制相关的过程，并通过源系统上的 SDR 开始 I/O 流。如果用户**终止** RCG，这不仅会停止站点之间的复制数据流，还会释放 SDR，不再代理 I/O 和写入日志。已终止或不活动的 RCG 不会占用额外的系统资源，只是一个配置占位符。

4.3 卷访问

目标卷不能设置为“Read and Write”。具有复制一致性组的目标卷的默认访问模式为“no access”。但在将目标卷映射到 SDC 时，用户可以选择将它们映射为“Read Only”。

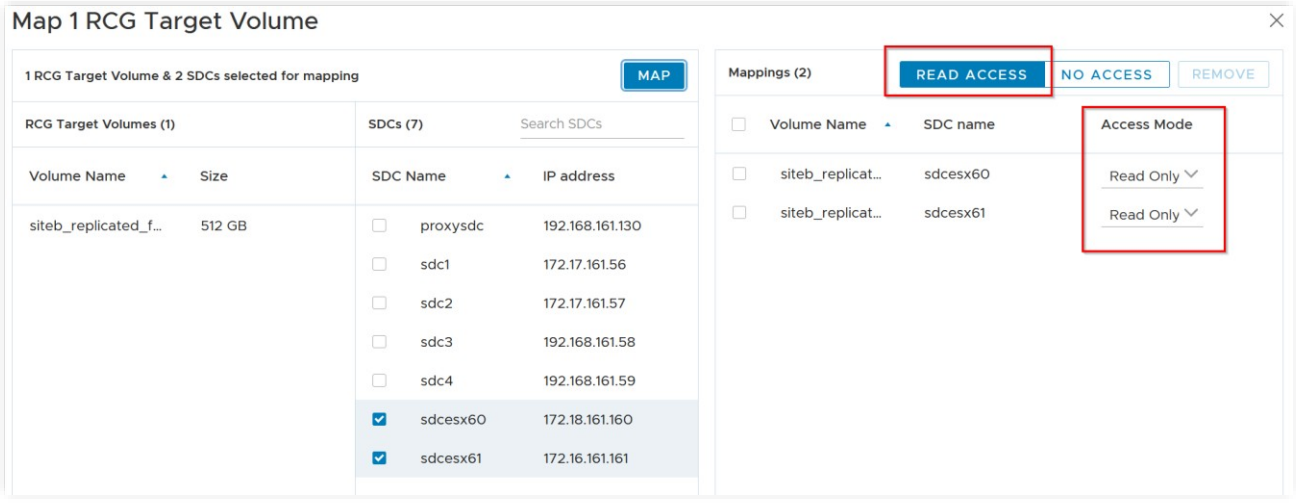


图 16

如果要在不创建快照的情况下检查目标卷上的数据，这会很有用。但是，除非还选择了“Freeze Apply”选项，否则该卷将继续从源接收更新的写入。

在配置复制的系统，以便让它也参与 VMware Site Recovery Manager 保护组时，必须将目标卷的只读属性映射到恢复站点 ESXi 主机。有关使用 VMware SRM 在采用复制 PowerFlex 卷的数据存储中保护虚拟机的更多信息，请参阅白皮书《使用 VMware Site Recovery Manager 对 Dell EMC PowerFlex 上的虚拟化工作负载进行灾难恢复》。

4.3.1 测试故障切换行为

PowerFlex 包括一个非常有用的工具，可以在不实际停止源端应用程序的情况下测试灾难恢复，并可故障切换到辅助站点。发出 **Test Failover** 命令的行为将会：

- 在目标系统上为连接到 RCG 的所有卷创建快照
- 将每个卷的卷映射所用的指针替换为指向其快照的指针
- 将目标系统 RCG 中的每个卷的快照/卷映射的访问模式设置为 **read_write**

这些步骤都在几毫秒内完成，如果卷之前映射为“只读”，这样就使 Storage Data Client 可立即写入卷。如果不这样，用户可以将目标卷映射到任何 SDC 以进行测试。由于您实际上是在映射快照，所以您可以对卷执行任何操作，无论是使用它们来打开数据库、应用程序，还是装载文件系统。由于它们是快照，因此您可以自由地测试您的应用程序，如果存储池与源系统具有相同的类型和组成，您的应用程序将同样表现出色。

在测试故障切换期间，源和目标之间暂停复制。但是，RCG 仍处于活动状态，因此写入仍流过 SDR，并在源端日志卷中累积。虽然用户可以使用测试故障切换功能对目标卷数据运行分析或执行其他测试操作，但使用下面讨论的创建快照功能可以更好地完成此类操作。

管理员可以使用测试故障切换功能安全地运行灾难恢复场景，而无需停机和维护时段。给出 **Test Failover Stop** 命令时，目标端指针会恢复为其原始状态，并再次指向复制卷本身。系统会删除快照，并会丢弃对它们进行的任何写入。最后，重新启动源和目标之间的数据复制，日志间隔将恢复传送。

4.3.2 故障切换行为

发出 RCG 故障切换命令时，原始源卷的访问模式将切换为 **read_only**。这意味着，在执行计划的故障切换时，您需要关闭应用程序。目标卷的访问模式切换为 **read_write**。不需要做别的事，如果从命令行界面或 REST API 发出故障切换命令，则行为也相同。如果计划了故障切换，但原始存储群集继续起作用，您可以选择启动 RCG 命令进行**反向**复制。这使卷对保持同步，只是现在方向相反。如果您要长时间关闭主系统，但希望保留 RCG 配置，则应终止 RCG 以将其设置为非活动状态。以后重新激活时，RCG 卷必须进行初始同步。

要记住的一件事是，每个 PowerFlex 存储系统都创建了唯一卷和 SCSI ID，因此源和目标系统的 ID 会不同。

4.3.3 创建快照行为

此 RCG 命令为连接到目标端 RCG 的所有卷创建快照，但不会进一步管理快照。使用快照是单独且手动的。在这里，要测试您的应用程序或使用其中包含的数据，您将需要：

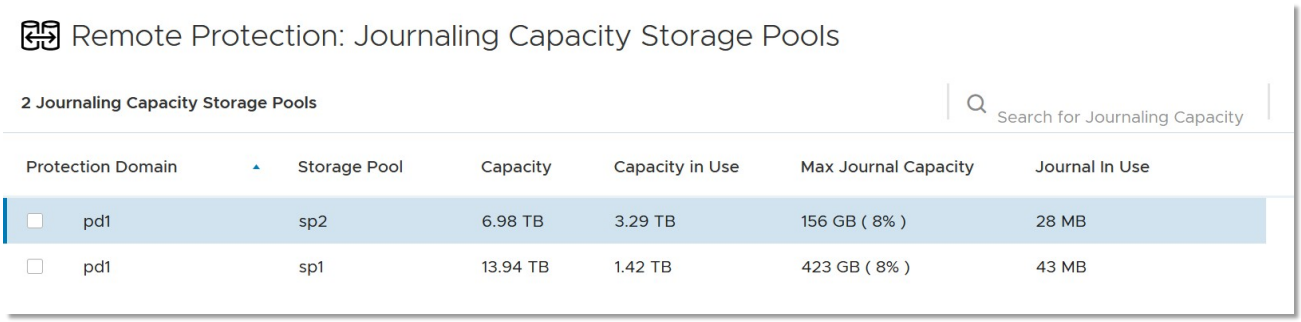
1. 将卷映射到目标 SDC 计算系统
2. 根据需要使用卷
3. 不再需要时取消这些卷的映射
4. 删除快照

我们在上面注意到，测试故障切换功能并不是很适合对目标端卷中的数据进行长时间运行或密集测试。但是，通过使用可写快照，用户可以：

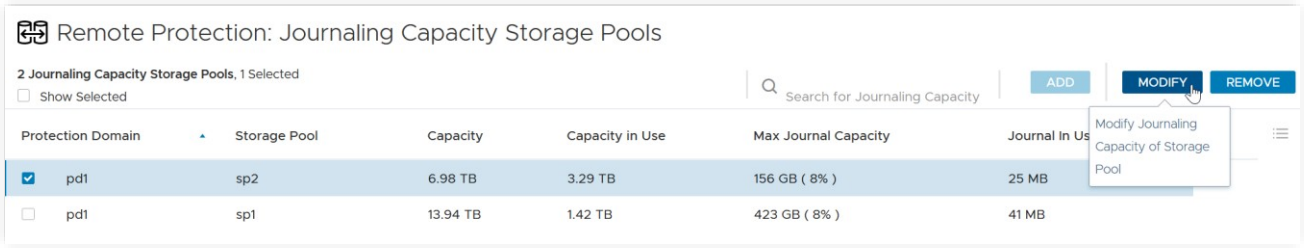
- 在辅助存储上执行资源密集型操作，而不会影响生产
- 在目标系统上测试应用程序升级，而不会影响生产
- 在目标环境中连接不同和更高性能的计算系统或介质
- 在目标域中连接具有不同硬件属性（例如 GPU）的系统
- 对数据运行分析，而不会妨碍您的操作系统
- 对数据执行“假设”操作，因为该数据不会写回到生产环境

4.3.4 监控日志容量和运行状况

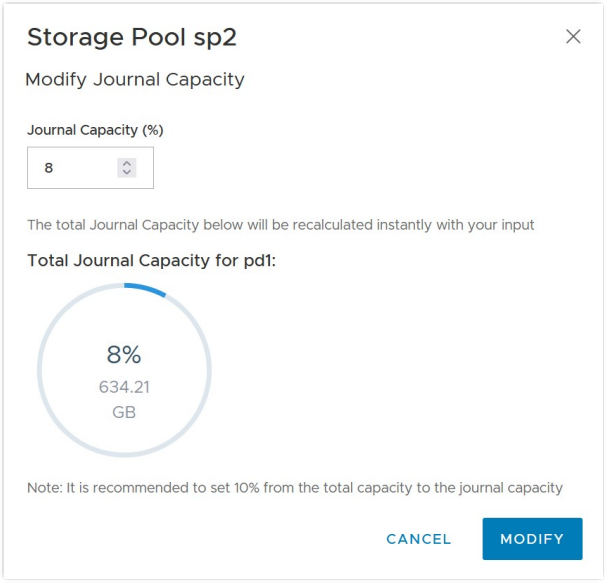
登录 WebUI 并导航到 PROTECTION → Remote → Journal Capacity，您可以跟踪日志空间预留的使用情况。



在这里，我们看到，我们为日志记录预留了 8% 或 156 GB 的存储池 sp2，而我们目前使用的日志容量只有 27 MB。如果担心预留的空间太小或太大，可以随时更改。选择存储池复选框，然后单击 MODIFY 命令。对预留进行任何必要的编辑。



我们在上面注意到，总体存储池容量的变化可能是增加或减少日志预留百分比的原因之一。另一个原因可能是要使用复制的卷或应用程序不断增加。



5 PowerFlex 复制网络注意事项

之前建议的所有网络拓扑、可用性和负载均衡选项仍然完全受支持。但是，PowerFlex 复制在网络构造中增加了一个必须考虑的新因素。与复制和相关的日志活动关联的额外网络开销。有关详细信息，请参阅《PowerFlex 网络最佳实践和设计注意事项》白皮书。

5.1 TCP/IP 端口注意事项

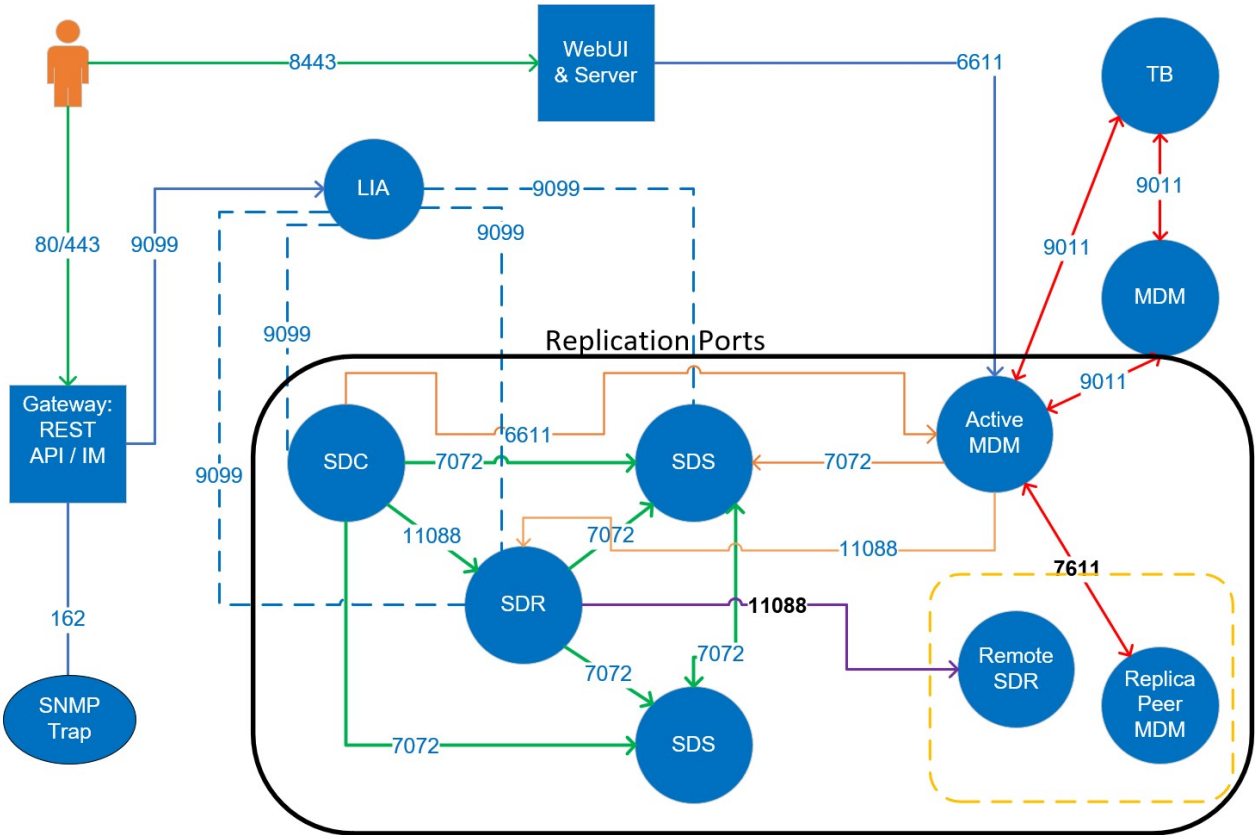


图 17 PowerFlex 端口和流量概述

上图为我们展示了 PowerFlex 的所有逻辑软件组件，以及这些组件所使用的 TCP/IP 端口。我们看到必须与 PowerFlex 服务器主机上的防火墙规则关联的端口。我们还可以看到专门与远程复制相关的端口，包括：

1. 将 SDC 和 MDM 链接到 SDR 的端口 11088 也会将 SDR 链接到远程 SDR。
2. 端口 7611 允许在两个复制群集之间进行 MDM 通信。

5.2 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。

5.3 网络带宽注意事项

首先，复制群集之间的通信有一些一般注意事项。**对复制卷的写入量不能超过群集之间的网络带宽。**计划好在群集之间的网络中至少有一条路径会失败，并确保可以维持预期的写入带宽，延迟在应用程序和服务级别的要求范围内。

5.3.1 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。

5.4 远程复制网络

对于通过静态路由 WAN 访问远程群集的网络配置，或网络基准延迟大于 50 ms 的网络配置，更值得关注的问题是延迟。**对于任何配置，都有 200 ms 的延迟限制**，这是地区性远程群集的潜在问题。对于超过 200 ms 的网络路径，从地球另一端接近的路径可能会表现得更好。此配置将需要至少两个连接到目标系统的子网，并且随着延迟的增加，带宽可能成为一个问题，因此，请彻底测试链路的延迟和吞吐量限制，并使复制带宽保持低于已知阈值。

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。通过 WAN 连接时，

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。

为 WAN 上的复制流量留出 20% 的安全裕度，考虑到应用程序 I/O 突发，以及考虑添加到 RCG 或重新激活的新卷的初始同步。

在某些情况下，当延迟较高时，您可能需要增大复制一致性组的 RPO。此操作可以在 RCGs 选项卡中完成。访问 **PROTECTION → Remote → RCGs**，选择一个 RCG，然后单击 **Modify→ Modify RPO** 命令来增大 RPO 值。

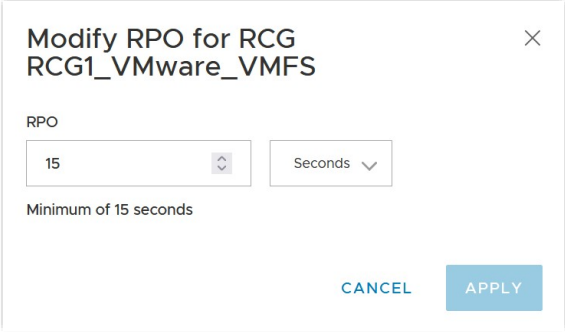


图 18

5.4.1 复制运行状况的网络影响

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。

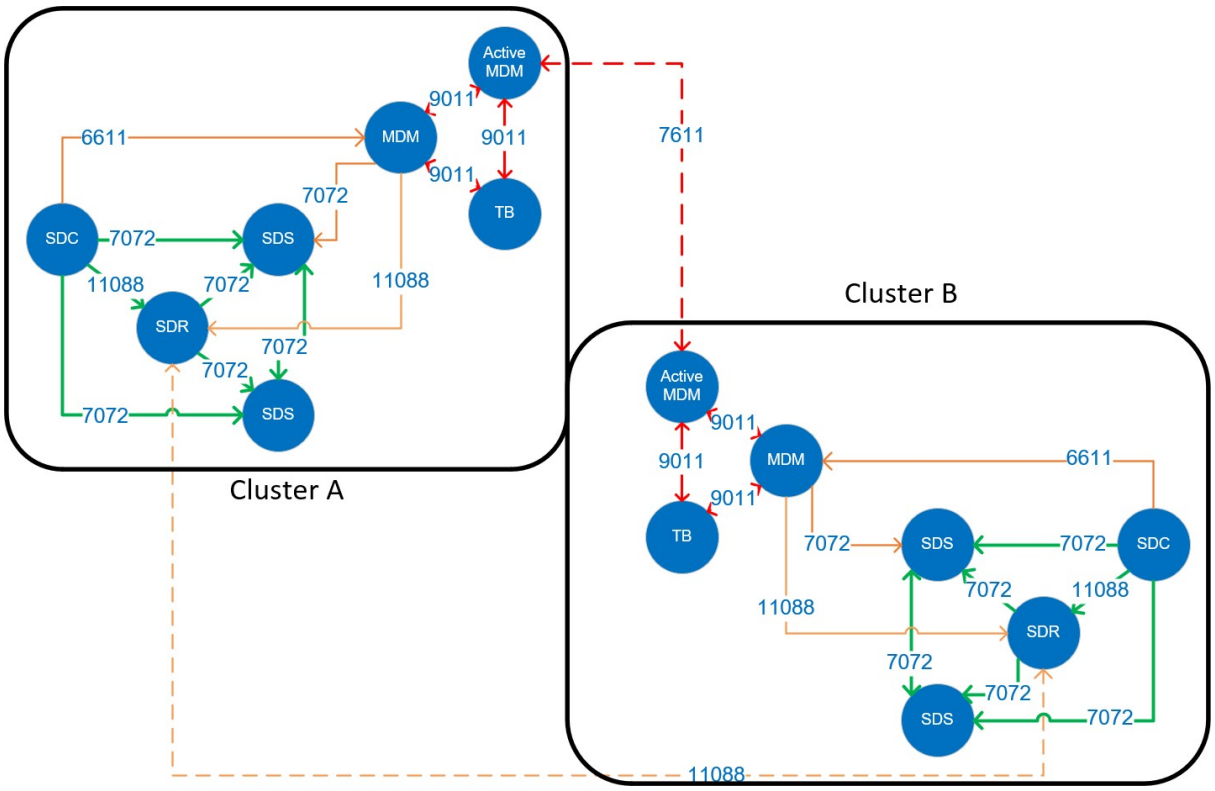
同样，日志卷容量应进行调整，以适应对等系统之间的链路中断。预计一个小时的停机可能是合理的，但我们建议用户规划 3 小时的停机。当链路中断时，RPO 明显会增大，必须确保有足够的日志空间来处理中断期间的写入。最好使用 PowerFlex Sizer 进行此类规划，但**通常情况下，日志容量应按如下进行计算：WAN 带宽 * 链路停机时间。**例如，如果 WAN 链路为 2x10 Gb（大约 2 GB/s），并且计划的停机时间为一小时，则日志大小应为 2 x 3600，或约为 7 TB。

当 WAN 链路恢复时，如果有 20% 的带宽预留空间，则系统可以赶上其最初的 RPO 目标。

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。

5.4.2 远程复制的路由和防火墙注意事项

第 5.1 节重点介绍了在复制群集之间的 MDM (7611) 通信，以及传输复制日志时使用的 SDR (11088) 通信的 TCP/IP 端口。



对于涉及远距离群集的复制应用场景，我们需要互连路由网络上提供的这些 IP 端口。在这种情况下，建立网络连接的最佳实践是为群集内 SDR 和 MDM 通信预留两个网络。

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。

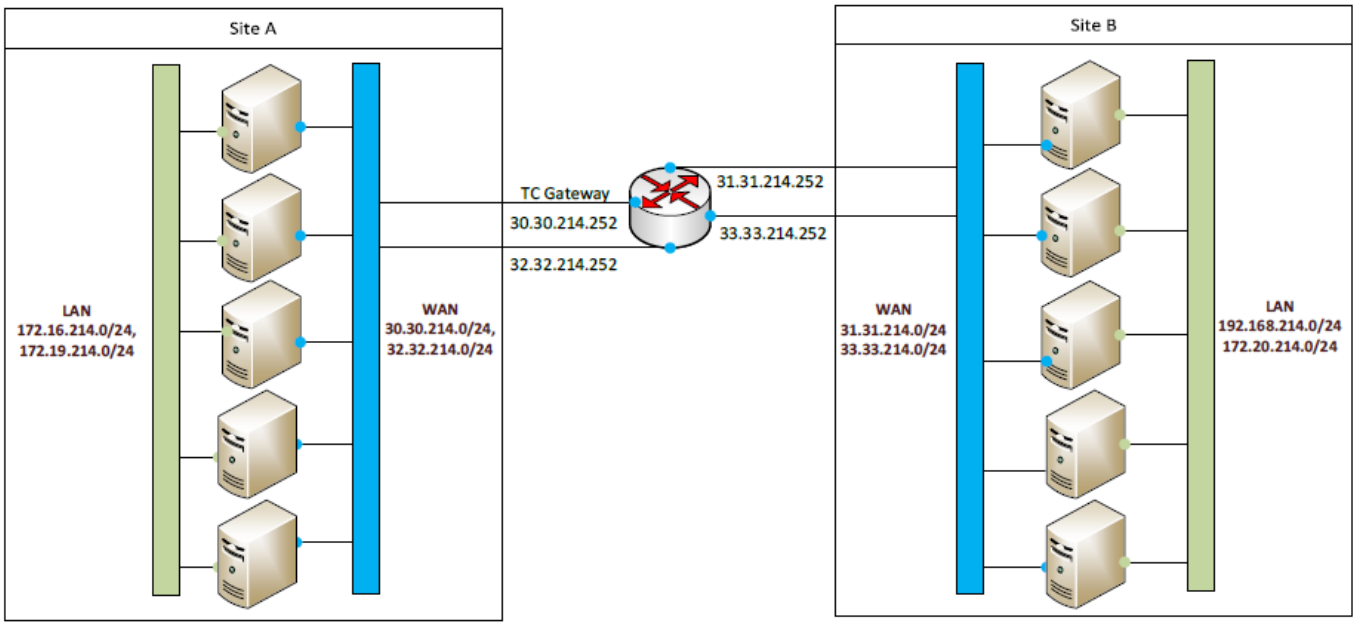


图 19　PowerFlex 复制的示例 WAN 拓扑。

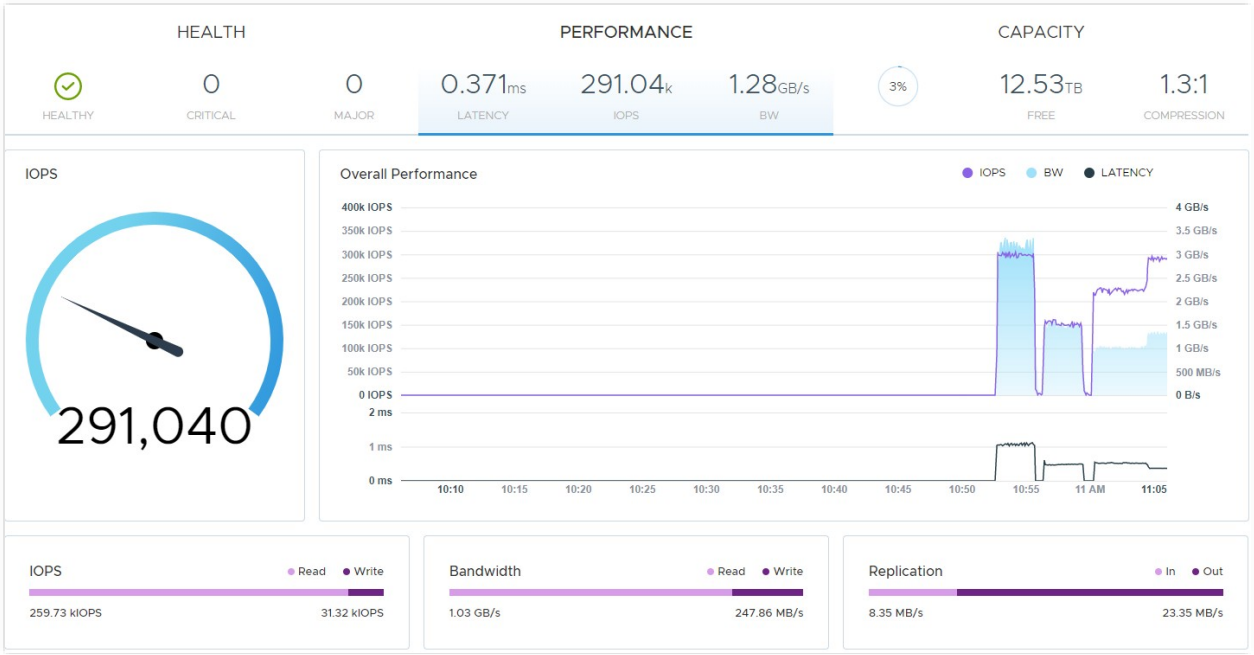
静态路由配置的详细信息将因操作系统/虚拟机管理程序和整体网络体系结构而异，但一般原则是相同的。

6 系统组件、网络和进程故障

对于与复制相关的最后一点考虑，我们必须面对服务器、进程和网络链接定期发生故障的现实。以下测试说明了其中一些类型的故障。在 6 节点 R740xd PowerFlex 节点群集上执行测试，每个存储池有三个 SSD。在发生故障时，两个存储池上的复制处于活动状态。

6.1 SDR 故障场景

我们从基准工作负载开始。我们将继续使 SDR 失效，观察其影响，并观察重新启动它的后续影响。



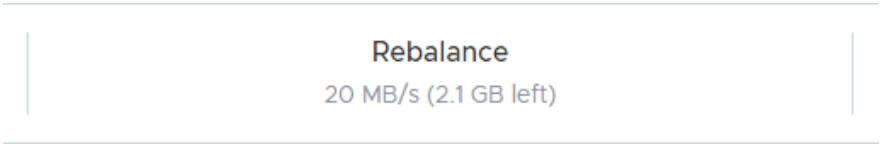
		
在使 SDR 失效后，我们可以立即看到 I/O 处理出现下降	I/O 恢复速度略慢一些	重新启动后，I/O 会受到轻微影响，但最终会回升到基准线。

6.2 SDS 故障场景

我们将对 SDS 故障执行相同的测试。

		
基准工作负载	SDS 失效后的 IOPs	五秒后，它开始回升，在 10 秒内恢复基准工作负载，表明该系统能够通过五个活动的 SDS 系统处理工作负载

正如预期的那样，我们看到重新平衡活动



	
接下来，我们重新启动出故障的 SDS，我们看到立即下降，但在重新启动 SDS 后并不显著	随着重新平衡继续，I/O 回升到基准线

6.3 网络链路故障场景

现在，我们将使一个网络链路失效，从而演示更新的原生负载均衡如何影响我们的 I/O 速率。系统的网络配置包括系统间的四个数据链路。

		
同样，我们确立了一个基准	我们使链路失效，并注意到 I/O 下降了 3 秒	5 秒后，恢复到基准

重新连接失效的端口后，基准 I/O 级别将在几秒内恢复，没有明显的下降。

所有这些故障场景都证明了 PowerFlex 的抗风险能力。它还表明系统经过了良好的调整，并且重建活动不会严重影响我们的工作负载。

7 复制 — 技术限制

下表列出了 PowerFlex 3.6 与复制相关的系统限制。

复制限制	
用于复制的目标系统数	1
每个系统的最大 SDR 数	128
复制一致性组 (RCG) 的最大数量	1024
RCG 中带有初始副本的最大复制对数	1024
每个 RCG 的最大卷对数	1024
每个系统的最大卷对数	32,000
远程保护域的最大数量	8
每个 RCG 的最大副本数	1
恢复点目标 (RPO)	最短：15 秒/最长：1 小时
最大复制卷大小	64 TB

8 总结

您现在应该对 PowerFlex 原生异步复制有了更好的了解，包括所选的配置和写日志方法。

总之，建议您从小规模开始。遵循上面提到的建议。考虑总体复制带宽，包括所有复制数据的所有写入 I/O。根据建议调整日志空间预留。包括网络和组件故障的误差幅度。