白皮书

ORACLE Platinum Partner

具备 EMC® VPLEX™ Metro 的远距离群集上的 Oracle Real Application Clusters (RAC) 最佳做法规划

摘要

本白皮书介绍与 Oracle Real Application Clusters (RAC) 和 Oracle Database 相关的 EMC® VPLEX™ 功能。其中还介绍了为以优化方式利用 EMC VPLEX Metro 而配置扩展 Oracle RAC 的最佳做法。

2012 年 5 月

EMC²

部件号 h8930.1

索引

执行摘要

EMC® VPLEX™ 是企业级存储联合技术，可在数据中心内部和之间聚合和管理光纤通道 (FC) 连接存储池。VPLEX 驻留在服务器和 FC 存储之间，并向主机呈现本地卷和分布式卷。VPLEX 存储聚合允许在线存储迁移和升级，且无需更改任何主机 LUN。VPLEX AccessAnywhere 群集技术可用于对分布式卷进行远程读/写访问，其中卷的 SCSI LUN 身份完全相同。这项技术让虚拟机管理程序可以远程迁移虚拟机 (VM)，并且简化扩展 Oracle Real Application Cluster (RAC) 跨数据中心的部署

VPLEX 产品线包括 VPLEX Local（单站点 SAN 联合）、VPLEX Metro（支持往返延迟高达 5 ms 的同步分布式卷）以及 VPLEX Geo（支持往返时间高达 50 ms 的异步分布式卷）。本白皮书重点介绍如何将 VPLEX Metro 与远距离群集中的 Oracle RAC（扩展 Oracle RAC）结合使用，以简化 SAN 基础架构，同时减少与扩展 RAC 部署模式关联的部分存储管理复杂性。

Oracle RAC 支持在服务器群集中透明地部署单个数据库，从而提供容错能力、高可用性和可扩展性。远距离群集中的 Oracle RAC 介绍的是一种部署模式，其中群集中的服务器驻留在物理分隔的位置。远距离群集中的 Oracle RAC 提供了一种横向扩展性能以及利用多个站点的存储和服务器资源的方法，而且与单站点 Oracle RAC 安装相比，还提高了整个站点故障情形或数据中心维护操作的恢复能力，且不会出现应用程序宕机。

VPLEX Metro 在扩展 Oracle RAC 中的主要好处：

- 在网络、存储和站点故障期间保证连续的数据库可用性
- 横向扩展体系结构以及两个站点（无空闲硬件）对同一数据库的完全读写访问权限
- 简化远程 Oracle RAC 部署：
    - 群集节点只需连接至本地 VPLEX 群集。不需要跨站点连接。
    - 通过使用分布式卷和 VPLEX Witness 部署 Oracle 表决磁盘，简化了基础架构要求。
    - 减少对 Oracle 群集节点 CPU 周期的占用以及与基于主机的镜像关联的占用。而是，由 VPLEX 提供的硬件 RAID 和底层物理阵列提供这项功能。
    - 能够创建可将多个数据库或应用程序文件作为一个单元进行保护的一致性组
    - 在存储硬件更新和迁移期间，VPLEX 卷不要求暂停应用程序或进行 LUN ID 更改

- 相对于单站点部署模式，扩展 Oracle RAC 模式可根据两个站点均可主动参与工作负载这个事实提供简易的灾难恢复测试和验证

EMC Symmetrix® VMAX™ 系列在本白皮书中用作 VPLEX 群集后面的存储阵列。Symmetrix VMAX 系列提供众多业界领先的特性和功能，可造就下一代高可用性虚拟数据中心和任务关键型应用程序。借助高级数据保护和复制，Symmetrix VMAX 系统处于企业存储区域网络 (SAN) 技术的最前沿。此外，采用 FAST VP 技术的 Symmetrix VMAX 系列以透明方式优化存储层来提高性能并节省成本，而且不会中断主机应用程序。

IT 组织正在寻求跨数据中心扩展数据库环境的方法，以便减少或避免与硬件故障、灾难、甚至是正常数据中心操作（如硬件刷新和迁移）关联的计划内和计划外宕机。借助 EMC VPLEX，这些组织将能更加灵活地改变其存储基础架构、灾难恢复能力以及增强协作和横向扩展体系结构。因此，EMC VPLEX 系统与 Oracle RAC 技术相结合，是这种新一代环境的理想选择。EMC VPLEX 的功能包括联合存储系统以及提供不受物理数据中心边界限制的全球独特设备，可与 Oracle RAC 的固有功能携手提供高可用性和可扩展的数据库访问。Oracle RAC 与 EMC VPLEX 和 Symmetrix VMAX 系列相结合是部署高度可靠的云计算环境的理想选择，可降低 IT 成本，同时提高基础架构效率。

目标读者

本白皮书面向 Oracle 数据库管理员、存储管理员和 IT 架构师，他们都负责设计、创建、管理和使用主要以 Oracle 数据库、VPLEX 技术和 Symmetrix VMAX 系列存储实现高可用性的 IT 环境。本白皮书假定读者对 Oracle RAC 和 Oracle 数据库技术、EMC VPLEX 和 Symmetrix 存储阵列有所了解。

简介

扩展 Oracle RAC 提供了一种横向扩展以及利用多个站点的存储和服务器资源的方法，而且提高了故障情形或维护操作的恢复能力，且不会出现应用程序宕机。这让企业可以消除数据库宕机，并实现无中断的持续业务处理，甚至在发生整个站点故障时也是如此。

请注意，尽管扩展群集部署模式中的 Oracle RAC 可跨单个数据库提供高可用性，但最好还是使用 RecoverPoint、SRDF®、Oracle Data Guard 或类似解决方案等技术跨更远距离部署灾难恢复 (DR) 解决方案。此类远程复制副本在发生数据库故障时很有帮助（例如，错误的文件或 LUN 删除、数据块损坏等） 此外，最佳做法是将备份策略部署到磁带或 VTL，其中可能会使用克隆/快照技术来减轻生产中的备份流程负担。

本白皮书向读者介绍了 EMC VPLEX 系列、VPLEX Metro 群集体系结构，以及与扩展 Oracle RAC 部署相关的特性和功能。本白皮书还讨论了扩展 Oracle RAC 解决方案在

各种故障情况下的恢复能力。其中还提供用于扩展 Oracle RAC 平台的 VPLEX 和 Symmetrix VMAX 系列存储的资源调配步骤，以及在采用 VPLEX Metro 和 Symmetrix VMAX 技术的 4 节点扩展 Oracle RAC 中运行 OLTP 工作负载的资源调配步骤。

产品和功能概述

VPLEX

EMC VPLEX 是适用于 EMC 和非 EMC 存储的存储虚拟化解决方案，如图 1 所示。VPLEX 后面的存储可以是异构式，既支持 EMC 存储，也支持 NetApp、HDS、HP 和 IBM 等其他存储供应商提供的常用阵列。

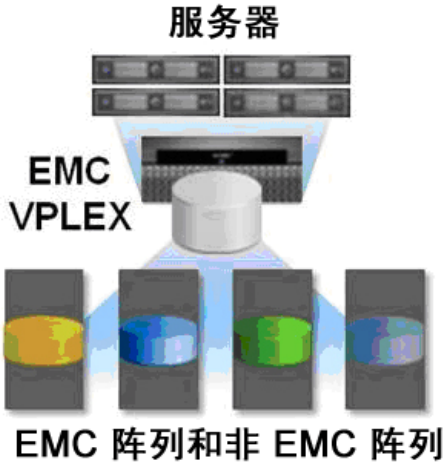


图 1. EMC VPLEX Local 用于联合异构存储

VPLEX 可以跨地理位置分散的数据中心扩展，以通过创建 VPLEX 分布式虚拟卷，提供存储设备的同时访问。VPLEX 技术可提供无中断的异构数据移动和卷管理功能。

由于这些功能，VPLEX 可提供具有差异优势的独特价值来满足三种不同的要求：

- 不管是在数据中心内部还是数据中心之间，都能跨不同计算和存储基础架构动态移动应用程序和数据。
- 可创建地理位置分散的高可用性存储和计算基础架构，而且恢复能力无与伦比。
- 可远程提供高效的实时数据协作。

VPLEX 产品服务

EMC 提供三种 VPLEX 配置，可满足客户对高可用性和数据移动性的需求，如图 2 所示：

- VPLEX Local
- VPLEX Metro
- VPLEX Geo（Oracle RAC 当前不支持）

图 2. VPLEX 拓扑

VPLEX Local

VPLEX Local 提供了无缝的无中断数据移动性，可让您从单一界面管理数据中心内的多个异构阵列。

VPLEX Local 可用于跨多个阵列提高可用性、简化管理以及提高利用率。

带 AccessAnywhere 的 VPLEX Metro

利用带 AccessAnywhere 的 VPLEX Metro，在往返时间 (RTT) 高达 5 ms 的同步距离内的两个站点之间实现主动-主动的数据块级别数据访问。

以下是两个使用 VPLEX Metro 和 Oracle 实现数据移动性和高可用性的示例。

- **应用程序和数据移动性** — 虚拟机管理程序本身可在物理服务器之间移动 VM，且没有应用程序宕机。与服务器虚拟化结合使用时，VPLEX 分布式卷允许用户以透明方式远距离移动和重定位 VM 及其对应的应用程序和数据。这提供了一项*独特功能*，让用户可以在站点之间重定位、共享和平衡基础架构资源。本白皮书将介绍采用 VPLEX Metro 的 Oracle VM 实时迁移移动性示例： 采用 EMC VPLEX 和 Symmetrix VMAX 系列实现 x86 实时迁移的 Oracle VM 服务器，如图 3 所示。

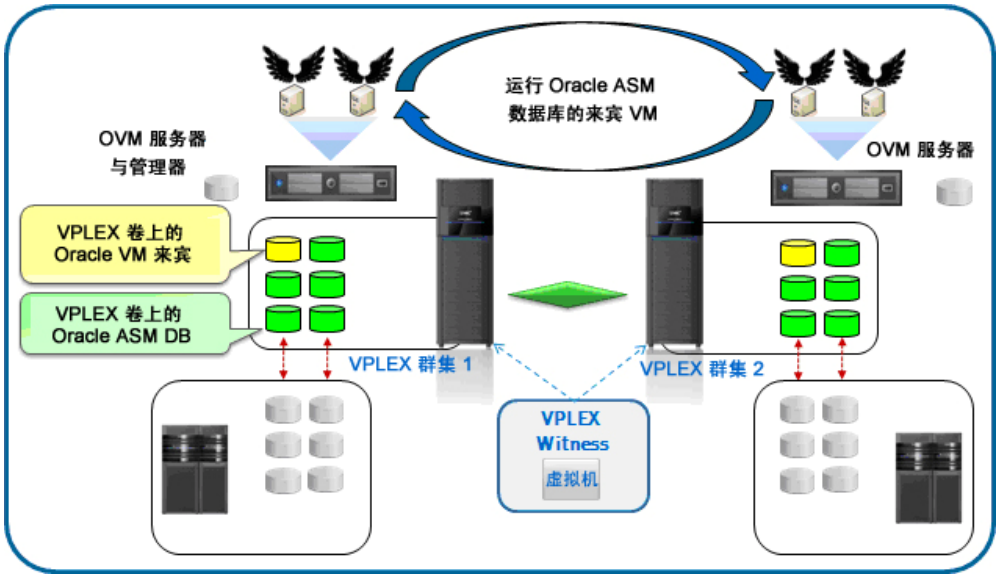


图 3. 采用 VPLEX Metro 的 Oracle VM 实时迁移

- **高可用性基础架构** — 缩短恢复时间目标 (RTO)。高可用性可为关键应用程序提供近乎持续的正常运行保障，并在发生故障时自动重启应用程序，同时尽可能减少人工干预。采用 VPLEX Metro 的扩展 Oracle RAC 示例如图 4 所示。在本白皮书将重点介绍此解决方案。

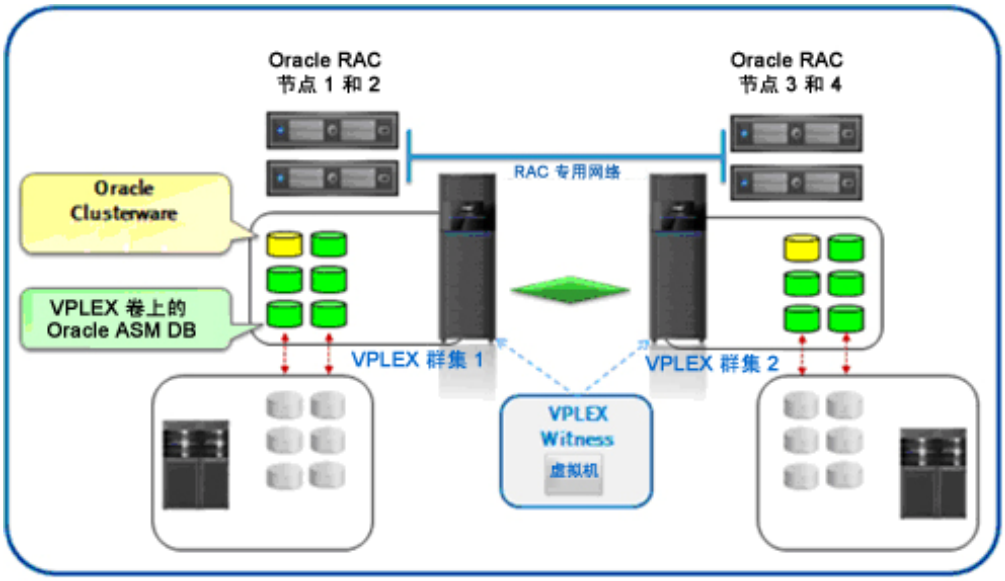


图 4. 采用 VPLEX Metro 的扩展 Oracle RAC

带 AccessAnywhere 的 VPLEX Geo

利用带 AccessAnywhere 的 VPLEX Geo，在往返时间 (RTT) 延迟高达 50ms 的异步距离内的两个站点之间实现主动-主动的数据块级别数据访问。*Oracle RAC 或扩展 Oracle RAC 当前不支持 VPLEX Geo，因此本文档不会详细介绍*。有关 VPLEX Geo 的详细信息，请访问 http://Powerlink.emc.com，关键字搜索： VPLEX Geo。

VPLEX 体系结构亮点[1]

VPLEX 系列使用一个独特的群集体系结构，可帮助客户打破数据中心的物理界限，并允许多个数据中心的服务器具有对共享数据块存储设备的读/写访问权限。VPLEX Local 包括单个群集，而 VPLEX Metro 包括两个。一个 VPLEX 群集包含一个、两个或四个引擎，如表 1 所示。每个 VPLEX 引擎都使用两个冗余控制器来提供 SAN/WAN 连接、缓存和处理能力。

表 1. VPLEX 硬件组件

功能	描述
VPLEX 群集	包含一个、两个或四个引擎
VPLEX 引擎	包含两个控制器、管理模块、电源、电池电源和风扇
VPLEX 控制器	包含一定数量 I/O 模块、SSD、CPU 和 RAM

VPLEX Local 使用直写缓存并允许直接传递写入，并且先获得 VPLEX 卷后面的存储的确认，再将确认发回给主机。借助 Symmetrix VMAX™ 系列和 VNX™ 系列等 EMC 存储（只需向存储持久缓存注册写入），可实现最佳的应用程序写入响应时间。VPLEX Metro 也使用直写缓存，但是，对应用程序的写入仅在向本地和远程存储注册后才会获得确认。在所有 VPLEX 部署中，读取可受益于 VPLEX 缓存，而且在 VPLEX Metro 中，从本地 VPLEX 群集缓存提供读取命中服务。

VPLEX 逻辑存储结构

VPLEX 封装了传统物理存储阵列设备，并将三层逻辑抽象应用于存储卷，如图 5 所示。扩展区是 VPLEX 用于划分存储卷的机制。扩展区可以是全部或部分底层存储卷。EMC VPLEX 聚合扩展区并可在设备层应用 RAID 保护。设备使用一个或多个扩展区构建，并可根据需要组合为更复杂的 RAID 方案和设备结构。VPLEX 存储结构的顶层是虚拟卷。虚拟卷通过设备创建，并且继承底层设备的大小。虚拟卷是 VPLEX 使用前端 (FE) 端口向主机公开的元素。虚拟卷访问使用存储视图控制，此类视图可与 EMC Symmetrix 上的自动资源调配组或 EMC CLARiiON® 上的存储组媲美。它们充当逻辑容器，可确定主机启动器对 VPLEX FE 端口和虚拟卷的访问。

[1] 本部分的详细信息基于 VPLEX 5.1 版，可能与其他版本存在差异。 VPLEX 产品指南提供了准确的版本详细信息。

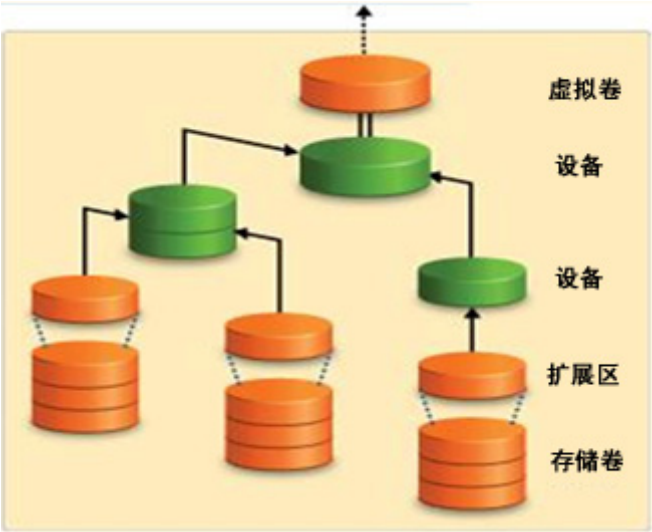


图 5. EMC VPLEX 逻辑存储结构

VPLEX 对存储区域网络 (SAN) 卷的封装方法是使用 WWN 标识存储设备，将设备打包为具备用户定义的配置和保护级别的 VPLEX 虚拟卷集，然后向主机提供虚拟卷。VPLEX 可以封装/解封现有存储设备（保留其数据），也可划分和聚合用于虚拟卷的现有存储卷。此外，通过 VPLEX 封装呈现给主机的虚拟存储也可无中断地在后端存储阵列内部和之间移动。建议在 VPLEX 封装 VMAX 系列和 VNX 系列等存储时，在存储阵列执行 RAID 保护。这将简化存储和 VPLEX 之间的映射，同时允许使用各项存储功能，如创建快照、克隆和附加 DR。

VPLEX 一致性组、分离规则和 Witness

一致性组

从 VPLEX Local 和 VPLEX Metro 的 GeoSynchrony 5.0 开始，一致性组就用于在出现站点丢失或 WAN 分区时，整理虚拟卷以及保证写入顺序保真度和确定性 I/O 连续行为。一致性组将卷聚合在一起，以向整个组提供公用属性集。此外，您还可以根据需要将一致性组从一个群集移至另一个群集。一致性组对于数据库和应用程序特别重要。所有数据库 LUN（例如，Oracle 数据、控制和日志文件）都需要保留写入顺序保真度，以维护数据完整性，因此总是应该一起放在单个一致性组中。通常，多个数据库具有事务相关性，如在将数据库链接用于连接数据库时，或在应用程序向多个数据库发出事务并期望它们互相一致时。在这些情况下，一致性组应该包括所有需要保留 IO 相关性（写入顺序保真度）的 LUN。

对于 Oracle RAC，仅支持 VPLEX Local 和 VPLEX Metro。Local 和 Metro 提供的唯一一致性组类型是同步（直写缓存模式）一致性组。同步一致性组的定义如下所示：

同步一致性组 — 提供一种将相同的 VPLEX 群集分离规则和其他属性应用于 VPLEX Local 或 VPLEX Metro 配置中的一组卷的简便方法，从而简化大型系统的系统配置和管理。同步组中的卷具有全局或本地可见性。同步一致性组使用直写缓存（在

VPLEX 用户界面称为同步缓存模式），而且在使用 VPLEX Metro 的情况中，在距离延迟高达 5 ms 的群集上受支持。这意味着，VPLEX Metro 将写入发送到后端存储卷，并在两个群集的后端存储卷都确认写入后立即确认对应用程序的写入。

分离规则

分离规则是在与远程群集的连接中断（例如，网络分区或远程群集故障）时，确定一致性组 I/O 处理语义的预定义规则。在这些情况下，在恢复通信之前，大多数工作负载需要特定虚拟卷集，才能在一个群集上继续 I/O 并在另一个群集上暂停 I/O。

在 VPLEX Metro 配置中，分离规则可以描述静态首选群集，方法是设置[2]：*winner:cluster-1*、*winner:cluster-2* 或 *No Automatic Winner（无自动优胜者）*（其中，最后一项指定无首选群集）。如果部署的系统没有 VPLEX Witness（将在下节中论述），一致性组设备 I/O 将在首选群集中继续，并在非首选群集中暂停。

VPLEX Witness

VPLEX Witness 随 GeoSynchrony 5.0 推出，是扩展 Oracle RAC 部署的必需组件。VPLEX Witness 通过管理 IP 网络连接至两个 VPLEX Metro 群集。VPLEX Witness 通过将其自身的观察与群集定期报告的信息进行协调，让群集可区分群集内网络分区故障和群集故障，并在这些情况下自动继续相应站点上的 I/O。VPLEX Witness 仅影响属于 VPLEX Metro 配置中同步一致性组成员的虚拟卷，并且仅当分离规则指明群集 1 或群集 2 是一致性组首选群集时才会影响（也就是说，“无自动优胜者”规则未生效时，VPLEX Witness 不影响一致性组）。

没有 VPLEX Witness 时，如果两个 VPLEX 群集失去联系，生效中的一致性组分离规则将定义哪个群集继续操作以及哪个暂停 I/O，如上所述。仅使用分离规则来控制哪个站点是优胜者时，可能会在出现站点故障时增加不必要的复杂性，因为可能需要手动干预才能恢复仍正常运行的站点 I/O。VPLEX Witness 会动态地自动处理此类事件，这也是它成为扩展 Oracle RAC 部署绝对必要项的原因。它提供了以下几项内容：

- 在数据中心之间自动实现负载平衡
- 主动/主动使用两个数据中心
- 存储层的完全自动故障处理

为了让 VPLEX Witness 能够正确区分各种故障情况，必须使用互不相同的网络接口在独立于任意群集的故障域中安装它。这将消除单个故障同时影响群集和 VPLEX Witness 的可能性。例如，如果将 VPLEX Metro 配置的两个群集部署在同一数据中心的两个不同楼层，请在不同楼层部署 VPLEX Witness。另一方面，如果将 VPLEX Metro 配置的两个群集部署在两个不同的数据中心，请在第三个数据中心部署 VPLEX Witness。

[2] 具体取决于管理 GUI 选项。 CLI 使用稍有差异的术语来指定相同的规则。

缓存存储区

为了在紧急情况下避免元数据丢失或数据丢失（使用写回缓存时），VPLEX 将使用称为缓存存储的机制来保护缓存到永久本地存储的信息。

Oracle Real Application Clusters

Oracle Real Application Cluster (RAC) 是屡获奖项的 Oracle Database Enterprise Edition 的一个选件。Oracle RAC 是采用共享缓存体系结构的群集数据库，可克服传统无共享和共享磁盘方法的限制，从而向所有业务应用程序提供高度可扩展且高度可用的数据库解决方案。Oracle RAC 支持所有种类的主流应用程序。这包括在线事务处理 (OLTP) 和决策支持系统 (DSS)。

Oracle Automatic Storage Management

Oracle Automatic Storage Management (ASM) 是用于 Oracle 数据库的集成文件系统和卷管理器。Oracle ASM 简化了数据库存储的管理。除了提供性能和可靠性优势之外，Oracle ASM 还可以提高数据库可用性，因为可以在线添加或删除 ASM 磁盘。在添加或删除磁盘后，Oracle ASM 可自动跨 ASM 磁盘组中的磁盘重新平衡数据。

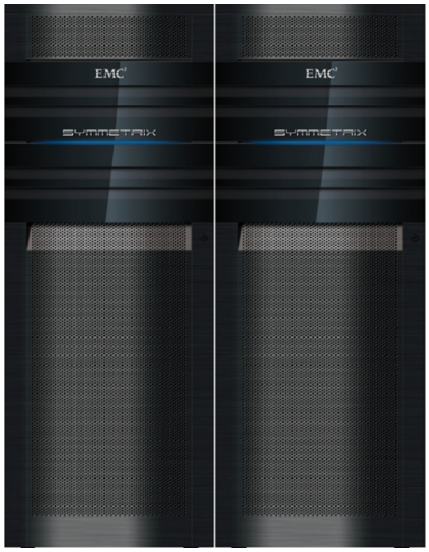
在 Oracle Database 11g 版本 2 中，Oracle ASM 和 Oracle Clusterware 已捆绑在名为 Oracle Grid Infrastructure 的软件包中。此软件包可提供运行 Oracle RAC 数据库所需的所有群集和存储管理服务。Oracle ASM 也经过了扩展，可支持将 Oracle 群集注册表 (OCR) 和表决文件放在 ASM 磁盘组内。

云距离群集中的 Oracle RAC

远距离群集中的 Oracle RAC 是一种部署模式，其中群集中的服务器驻留在物理分隔的位置。远距离群集中 Oracle RAC 提供的可用性更胜于本地 Oracle RAC。远距离群集中的 Oracle RAC 可提供站点故障的极速恢复，并且允许所有站点中的所有服务器作为单个数据库群集的一部分主动处理事务。即使此体系结构令您很感兴趣并已成功实施，也务必了解此体系结构最适合的情况，尤其是距离、延迟和提供的保护度等方面。延迟及距离产生的重大影响将会对此体系结构的部署位置造成一些实际限制。此体系结构最适合两个数据中心相对较近的情况，以及已支付在站点之间设置直接专用通道的成本的情况。

Symmetrix VMAX™ 系列

Symmetrix VMAX 系列以简单、智能、模块化存储这一策略为基础，整合了新的 Virtual Matrix™ 互连；此互连可跨所有节点连接和共享资源，从而使存储阵列可以从入门级配置无缝地扩展为世界上最大的存储系统。它可提供最高级别的性能和可用性，并具备一些新的硬件功能，如图 6 所示。



- 2 – 16个控制器板
- 高达 2.1 PB 的可用容量
- 高达 128 个 FC FE 端口
- 高达 64 个 FICON FE 端口
- 高达 64 个 GigE / iSCSI FE 端口
- 高达 1 TB 的全局内存（512 GB 可用）

图 6. Symmetrix VMAX 系列

Symmetrix® VMAX™ 系列提供终极横向扩展平台。它包含通过增加处理模块（节点）和存储机架以增量方式扩展前端和后端性能的功能。每个处理模块都提供了额外的前端、内存和后端连接。

Symmetrix® VMAX™ 系列还将最大超级卷大小增加到 240 GB（在 Symmetrix DMX™ 上为 64 GB）。这样可以轻松地进行存储规划和设备分配，特别是在使用 Virtual Provisioning™ 时（精简存储池已经分条，可以轻松使用大型超级卷）更是如此。

Symmetrix VMAX™ TimeFinder 产品系列

利用 EMC TimeFinder® 本地复制技术系列，可以为数据库和应用程序数据创建多个基于存储的无中断、可读/写复制副本。它具有高速、可扩展、存储利用效率高以及对应用程序的影响最小乃至无影响的特点，且这些特点不受数据库大小影响，因而可满足众多种客户的数据复制需求。TimeFinder 提供了用于备份、重启和恢复生产数据库和应用程序的解决方案，即使这些数据库和应用程序跨多个 Symmetrix VMAX 也不受影响。TimeFinder 与其他 EMC 产品（如 SRDF）完美集成，利用它，无需中断同步或异步复制即可创建远程目标的副本。如果需要用远程副本进行还原，则 TimeFinder 和 SRDF 将以增量方式并行地还原数据，以便实现最高级别的可用性和保护。TimeFinder 产品系列支持利用 EMC 一致性技术创建相关的写入一致性复制副本，并支持创建对 Oracle 备份/恢复操作有效的复制副本，后文的使用情形中对此进行了说明。除了 Timefinder，Symmetrix VMAX™ 系列还具备集成 RecoverPoint 写拆分器，可提供 RecoverPoint™ 本地和远程数据保护技术。有关 RecoverPoint 3.5 与 Symmetrix VMAX 系列集成的详细信息，请参阅 http://china.emc.com 和 http://Powerlink.EMC.com。

Symmetrix VMAX Virtual Provisioning

Symmetrix 精简设备是逻辑设备，可以按照 Symmetrix 设备的众多传统使用方式来使用这些设备。与传统 Symmetrix 设备不同，在创建精简设备并将其呈现给主机时，无需为其预分配物理存储（尽管在某些情况下，只对精简池宽分条和管理简易性感兴趣的客户会选择完全预分配精简设备）。您将精简设备绑定至精简池之前，无法使用此精简设备。多个精简设备可以与任意给定的精简池绑定。精简池由称为数据设备的设备组成，这些设备提供实际物理存储来支持精简设备分配。表 2 介绍了基本虚拟资源调配定义。

表 2. 虚拟资源调配设备的定义

设备	描述
精简设备	主机可访问的未直接关联存储的设备。
数据设备	放在精简池中可提供精简设备要使用的存储容量的内部设备。
精简池	为精简设备提供存储容量的数据设备的集合。

在扩展 Oracle RAC 中实施 VPLEX Metro

VPLEX Metro 环境中的扩展 Oracle RAC 部署注意事项

EMC VPLEX 打破了数据中心的物理壁垒，允许用户并发访问位于不同地理位置的数据。采用 VPLEX Metro 的扩展 Oracle RAC 允许在访问单个数据库的多个站点之间透明地共享工作负载，同时可以在预见到计划内事件（例如硬件维护）时灵活地迁移站点间的工作负载。此外，如果发生的计划外事件导致其中一个数据中心中断服务，则可使用 Oracle Transparent Application Failover (TAF) 自动地将出现故障的客户端连接重定向到仍正常运行的站点上运行的 Oracle RAC 节点。

扩展 Oracle RAC 和 VPLEX Metro 部署指导准则

以下几点介绍一部分主要部署指导准则。

Oracle Clusterware 和 VPLEX Witness 部署

通常，扩展 Oracle RAC 的部署重点介绍为其中一个 Oracle Clusterware 表决文件部署第三个站点（可能基于 NFS）。明确地说，在采用 VPLEX Metro 的远距离群集中，Oracle RAC 仍需使用 Oracle Clusterware 表决磁盘。但是，群集表决磁盘本身驻留在 VPLEX 虚拟卷上。这可保证 Oracle 表决磁盘访问/Oracle RAC 行为和 VPLEX Metro 故障切换行为一致。借助 VPLEX，仅 VPLEX Witness 部署在独立故障域中（多站点部署中的第三个站点），如前面的 VPLEX Witness 部分所述：

- 如果只有 Oracle 互连分区（不是真正的站点故障，且不影响 VPLEX 互连），则 Oracle Clusterware 将根据多数节点及表决磁盘访问进行重新配置。

- 如果是 VPLEX 互连分区（或真正的站点故障），VPLEX 将根据站点首选项规则和 Cluster Witness 指导，立即允许 IO 在一个群集恢复。因此，仅当 VPLEX 恢复表决磁盘上的 I/O 时，这些 Oracle 群集节点才有权访问这些表决磁盘；而且 Oracle Clusterware 将相应地重新配置群集。尽管仍需要表决磁盘，但无需部署在独立的第 3 个站点中，因为 VPLEX Witness 可提供 Split-Brain 保护并保证 Metro 和 Oracle Clusterware 的行为一致。此外，由于 VPLEX Witness 控制表决文件的访问，因此可跨独立 Oracle RAC 部署和相关上游用户应用程序保证一致的确定性行为。

Oracle Clusterware 部署

Oracle Clusterware 仅部署在 VPLEX 分布式卷上（而不部署在第三个站点），如上节所述。在 Oracle Database 11g 版本 2 中，Oracle Clusterware 与 Oracle ASM 合并创建了 Oracle Grid Infrastructure。因此，将在安装 Oracle Grid Infrastructure 时创建第一个 ASM 磁盘组。

- 将 Oracle ASM 用于托管 Oracle Clusterware 文件（OCR 和表决文件）（从 Oracle Database 11g 版本 2 开始）时，EMC 建议仅为 Oracle Clusterware 文件创建唯一磁盘组，例如： +GRID（也就是说，不在此处放置任何数据库内容，如日志或数据文件）。+GRID 磁盘组将从使用正常冗余或高冗余中受益。这样，Oracle 将创建多个表决磁盘（而不是像 +GRID ASM 磁盘组使用外部冗余那样仅创建一个[3]）。由于未包括数据库内容，因此用于此磁盘组的分布式 VPLEX 设备大小会相对非常小。
- 在 VPLEX 后面使用 EMC 存储（例如，Symmetrix VMAX 系列或 VNX 系列）时，或在使用 VPLEX RAID 保护时，建议方法是将其他所有 ASM 磁盘组设置为外部冗余。这将根据 VPLEX 或 EMC 存储阵列 RAID，向 ASM 成员提供足够的保护。
- 由于所有 Oracle 群集节点都需要访问所有 Oracle Clusterware 和数据库设备，因此 Oracle Clusterware 和 Oracle 数据库都只能使用 VPLEX 分布式卷，而不管是使用 ASM、原始设备还是群集文件系统。

附加说明

在基于 x86 的服务器平台上，确保分区对齐。VPLEX 需要以 4 KB 偏移量对齐；但是，如果使用 Symmetrix，请以 64 KB（128 个数据块）偏移量对齐（本机也以 4 KB 边界对齐）：

- 在 Windows 上，可以使用 diskpar 或 diskpart。在 Linux 上，可以使用 fdisk 或 parted。
- 在本节后文中将会显示使用 fdisk 以 64 KB 偏移量对齐分区的示例： 在 PowerPath 设备上创建分区.

[3] 从 Oracle 11g 版本 2 开始，Oracle Clusterware 表决文件数量由 ASM 冗余级别自动确定。 例如，外部冗余表示 1 个表决磁盘，正常冗余表示 3 个表决磁盘，而高冗余表示 5 个表决磁盘。

扩展 Oracle RAC 和 VPLEX Metro 可防止计划外宕机

将扩展 Oracle RAC 和 VPLEX Metro 相结合时，可在许多故障情况下提高可用性和恢复能力，进而提高任务关键型数据库和应用程序的可用性。做法的详细信息。

表 3汇总了故障情形以及将让数据库能够在每种情形下继续操作的最佳做法的列表。请注意，该列表不涵盖如何故障切换到备用系统（如 Oracle Data Guard、RecoverPoint、SRDF 等）。《EMC VPLEX with GeoSynchrony 5.0.1 and 5.1 Product Guides》（带 GeoSynchrony 5.0.1 和 5.1 的 EMC VPLEX 产品指南）提供了有关 VPLEX 连接最佳做法的详细信息。

表 3. VPLEX Metro、Oracle RAC 和扩展 Oracle RAC 恢复能力的摘要

主机和站点故障情形的恢复能力			
Failure	Oracle 数据库单服务器（非 RAC）	Oracle RAC（未扩展）	采用 VPLEX Metro 的扩展 Oracle RAC
主机 HBA 端口故障	• 每个主机都应该具有多条存储路径。使用多个 HBA 端口（启动器）。 • 使用多路径软件（如 EMC PowerPath®）实现自动路径故障切换和负载平衡。 • 理想 SAN 连接将使用冗余交换机，其中 HBA 端口（启动器）将跨这些交换机分布。	与单服务器相同	与单服务器相同
主机硬件故障或崩溃	• 隐含的宕机将持续至主机和应用程序可以恢复操作。	• Oracle RAC 可针对 N-1 个节点的故障提供数据库恢复能力（其中 N = 群集中的节点数），方法是执行自动实例恢复并让其他群集节点随时可供用户连接使用。 • Oracle Transparent Application Failover (ATF) 可用于允许会话自动故障切换到仍正常运行的 RAC 群集节点。	与 Oracle RAC 相同

实验室/建筑/站点故障	• 隐含的宕机将持续至主机和应用程序可以恢复操作。	• 隐含的宕机将持续至主机和应用程序可以恢复操作。	• 通过在独立故障域（如其他建筑或站点）中安装 VPLEX 群集和 Witness，可在实验室、建筑或站点故障中恢复。 • 故障域中不受灾难影响的 VPLEX 群集将继续服务于应用程序 I/O。 • 使用 Oracle Transparent Application Failover 允许用户连接自动故障切换到仍正常运行的群集节点。
数据库/网络相关故障情形的恢复能力			
Failure	Oracle 数据库单服务器（非 RAC）	Oracle RAC（未扩展）	采用 VPLEX Metro 的扩展 Oracle RAC
数据库实例崩溃或公用网络断开连接	• 隐含的宕机将持续至实例可以重新启动或公用网络重新连接。	• Oracle RAC 可针对 N-1 个节点的故障提供数据库恢复能力（其中 N = 群集中的节点数），方法是执行自动实例恢复并让其他群集节点随时可供用户连接使用。 • Oracle Transparent Application Failover 可用于允许会话自动故障切换到仍正常运行的群集节点。	与 Oracle RAC 相同
Oracle RAC 互连分区	• 不适用	• Oracle RAC 可根据重新配置的群集，在本机自动处理此故障情形。	与 Oracle RAC 相同

存储故障情形的恢复能力			
Failure	Oracle 数据库单服务器（非 RAC）	Oracle RAC（未扩展/扩展）	采用 VPLEX Metro 的扩展 Oracle RAC
前端端口故障	• SAN 连接应该包括多个存储前端端口，理想情况下是跨多个 Symmetrix 控制器。如果使用的是具备多个引擎的 Symmetrix，请也连接到不同引擎上的端口，以获取更高的保护。	与单服务器相同	与单服务器相同
物理驱动器故障	• 使用存储 RAID 保护。Symmetrix 存储使用 RAID 保护，其中 RAID1 和 RAID5 可防止 RAID 组内的单磁盘故障，而 RAID6 可防止 RAID 组内的双磁盘故障。在任一情况下，应用程序都将在不受干扰的情况下继续运行。 • 如果驱动器启动失败，Symmetrix 热备盘驱动器将拷贝其数据，而 EMC Enginuity™ 将立即启动 Call-Home 以通知 EMC 支持部门。	与单服务器相同	与单服务器相同
存储阵列组件，包括控制器板（缓存，IO）	• Symmetrix 组件完全冗余，还包括持久镜像缓存（在电源故障延长时使用存储）、冗余控制器和电源。 • Symmetrix 数据从进入存储到离开这段时间内受到 T10 DIF 的保护。	与单服务器相同	与单服务器相同
与存储阵列的连接中断	• 隐含的宕机将持续至可以恢复存储阵列连接。	• 隐含的宕机将持续至跨阵列配置基于主机 (ASM) 的镜像，或者持续至可以恢复存储阵列连接。	• VPLEX Metro 同步一致性组将继续服务于两个站点的 I/O，即使其中一个存储阵列不可用也是如此。 • Oracle Clusterware 不会察觉到存储不可用，因为 VPLEX 群集会继续服务于所有 I/O。

VPLEX 故障情形的恢复能力			
Failure	Oracle 数据库单服务器（非 RAC）	Oracle RAC（未扩展/扩展）	采用 VPLEX Metro 的扩展 Oracle RAC
前端端口	• SAN 连接应该包括多个存储前端端口，理想情况下是跨多个 Symmetrix 控制器。如果使用的是具备多个引擎的 VPLEX，请也连接到不同引擎上的端口，以获取更高的保护。 • 使用多路径软件（如 PowerPath）实现自动路径故障切换和负载平衡。 • 理想 SAN 连接将使用已连接到多个 VPLEX 前端端口的冗余交换机。	与单服务器相同	与单服务器相同
后端端口	• 与 VPLEX 前端端口方法类似，也使用与 VPLEX 后端端口的冗余交换机连接，因为它们连接至存储阵列。	与单服务器相同	与单服务器相同
VPLEX 硬件组件	• VPLEX 组件完全冗余，包括持久缓存（在电源故障延长时使用存储）、冗余控制器和电源。	与单服务器相同	与单服务器相同
VPLEX 互连分区	不适用	不适用	如果两个站点仍可用，VPLEX 首选群集分离规则将确定哪个群集恢复 I/O 以及哪个暂停，而且连接到仍正常运行的群集的主机不会出现宕机。

VPLEX 群集不可用	• 不适用	• 不适用	• VPLEX Witness 将允许在仍正常运行的 VPLEX 群集上恢复 I/O。连接到该 VPLEX 群集的 Oracle RAC 节点将继续操作，且 RTO 为 0。 • 使用 Oracle Transparent Application Failover (TAF) 允许客户端自动与连接到仍正常运行的 VPLEX 群集的 Oracle RAC 节点重新建立连接

尽管 VPLEX Metro 的体系结构旨在支持多个位置的并发访问，但是产品的当前版本支持按两个站点之间的最大往返延迟为 5 ms 的同步距离分隔的双站点配置。此外，采用 VPLEX Metro 的扩展 Oracle RAC 可能需要将 VLAN 扩展到不同物理数据中心，以方便客户端连接和 Oracle RAC 节点互连。可以利用诸如 Brocade VPLS 和 Cisco Overlay Transport Virtualization (OTV) 等技术提供此服务。《EMC VPLEX Architecture and Deployment: Enabling the Journey to the Private Cloud》（EMC VPLEX 体系结构和部署：支持私有云之旅）技术书籍（位于 EMC Powerlink® 上）提供了有关 EMC VPLEX Metro 配置的详细信息。

扩展 Oracle RAC 实验室配置和测试中的 VPLEX Metro

实验室配置和设置

以下部分介绍用于本白皮书中记录的测试情形的技术和组件。

物理环境

图 7 演示用于本白皮书中所示测试的扩展 Oracle RAC 配置部署的总体物理体系结构。扩展 Oracle RAC 包含四个 Oracle RAC 节点，每个模拟数据中心（站点 A 和 站点 B）各有两个节点。

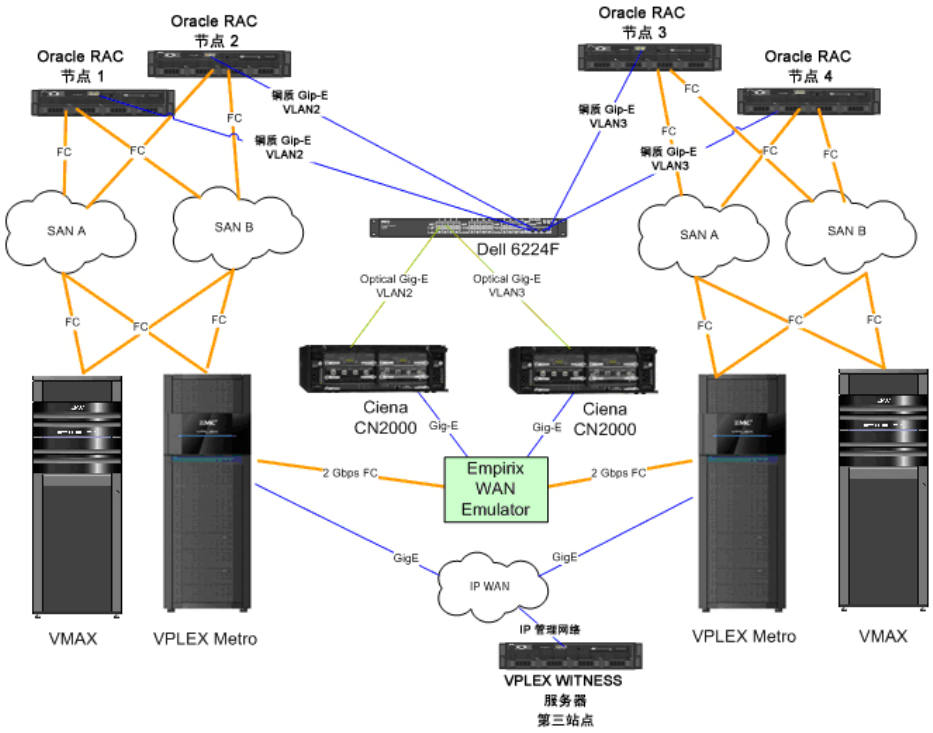


图 7. 扩展 Oracle RAC 的 VPLEX Metro 配置

硬件设置包括在主机、VPLEX 群集和 Symmetrix VMAX 系列存储中设置与冗余交换机的光纤通道连接。在每个站点中，主机将分区到其本地 VPLEX 群集前端端口，而存储将分区到 VPLEX 群集后端端口。VPLEX 群集将彼此分区（在此案例中是通过距离模拟器实现的）。站点间的 Oracle RAC 互连使用相同的距离模拟器体验与 VPLEX 互连相同的延迟开销。距离模拟器使用 2 个具备 WAN 压缩的 1 GigE 链路和 Ciena 交换机来支持协议更改。

表 4 显示了 VPLEX、主机和 Symmetrix 存储阵列硬件详细信息。它们包括 Metro 配置中的 2 个双引擎 VPLEX 群集设置以及 2 个单引擎 Symmetrix VMAX 系列技术，从而向每个 VPLEX 群集提供本地存储。4 个 RAC 节点包含 Dell 2950 双核服务器，其中缓存为 16 GB。请注意，测试重点是使用 Oracle Automated Stress Tool (OAST) 生成高 IO 工作负载，而不是使用大量 RAM 来实现高事务率数字。VPLEX Witness 针对每个 VPLEX 群集进行了设置且与其建立了连接。

表 5 显示距离模拟硬件。它包括的 Dell PowerConnect 6224F 光纤以太网交换机可将千兆位 IP 网络转换为光纤通道网络。Ciena CN 2000 Network Manager 提供具备光纤通道、FICON 或 GbE 输入的虚拟光纤网络。它同时支持数据压缩和动态带宽分配 (DBA)，以向客户端连接提供端口级服务质量。Empirix PacketSphere Network Emulator 针对诸如网络延迟之类的 IP 网络故障提供受控网络模拟，以实现全双工千兆位线速网络流量。

表 4. 扩展 Oracle RAC 硬件环境

硬件	数量	版本和配置
EMC VPLEX Metro	2	带 GeoSynchrony 5.0.1 的 VPLEX Metro 每个群集上有两个引擎和四个控制器
VPLEX Witness	1	运行 VPLEX Witness 虚拟机 (VM) 的 Dell R900
Symmetrix VMAX 系列	2	带 Enginuity 5875 的单引擎 VMAX，112 个使用虚拟资源调配的 450 GB/15k FC 驱动器
Dell 2950 服务器（RAC 节点）	4	2 个双核 16 GB RAM
Emulex Express（每台服务器有 2 个 HBA 端口）	4	

连接步骤中的主要指导准则是最大程度提高硬件冗余，例如使用两个交换机、多个 HBA、多路径来实现动态路径故障切换和负载平衡。

表 5. 距离模拟硬件环境

硬件	数量	版本和配置
Dell PowerConnect 6224F 光纤以太网交换机	1	最多支持 4 条 10 Gb 光纤和 2 条 10GBase-T 铜质以太网上行链路
Ciena CN 2000 Network Manager	1	Ciena ON-Center CN 2000 Network Manager 5.0.1
Empirix PacketSphere Network Emulator	1	Empirix Network Emulator NE XG 1.0B8，2 个 1-GbE 网络模拟器
Dell 2950 服务器	4	2 个双核 16 GB RAM

表 6 介绍了使用的主机软件。

表 6. 主机软件

软件	发布
服务器 OS	Oracle Linux 版本 5 更新 4 x86_64
EMC PowerPath	Linux x86_64 版本 5.5
Oracle	用于 Linux x86-64 的 Oracle Clusterware 11g R2 (11.2.02) 和 Oracle Database 11g R2 (11.2.0.2)

存储设置和设备分配规划

Symmetrix VMAX Virtual Provisioning 和存储设备配置

表 7 介绍了带存储虚拟资源调配的 Symmetrix VMAX 设备配置以及具备扩展 Oracle RAC 测试环境的 VPLEX Metro 的卷布局。

此配置将 Oracle 数据和日志文件放在单独的精简池中，允许每个文件都使用不同的 RAID 保护。在此配置中，数据文件放在 RAID5 保护的精简池中，而重做日志放在 RAID1 保护的精简池中：

- Symmetrix RAID5 保护的精简池为数据文件提供了保护、性能和容量利用率好处的绝佳组合，而且写入和旋转奇偶校验得到了优化，因此是数据文件的理想选择。在某些情况下，RAID1 保护的精简池可提供略胜于 RAID5 的可用性和性能，因此用于日志文件。请注意，两个精简池都共享同一组物理磁盘，以便允许完全共享物理资源。
- 值得注意的是，当数据文件和日志文件共享相同的精简池和 RAID 保护（RAID1、RAID5 或 RAID6）时，可以部署注重简单性（而非纯存储性能/可用性优化）的不同配置。

在配置中使用多个 ASM 磁盘组：

- +GRID： 如前面的 Oracle Clusterware 部署部分所述，将 ASM 用于 Oracle Clusterware（从 Oracle Database 11g 版本 2 开始）时，EMC 建议仅为 CRS 创建唯一磁盘组，例如： +GRID。在计划使用克隆或快照等存储技术为重新调整用途、备份等创建附加数据库拷贝时，这种将 Clusterware LUN 与数据库分开的方法很有用。这些拷贝不包括 Oracle Clusterware LUN。这对 RecoverPoint 和 SRDF 等 DR 解决方案也很有帮助，因为复制的 ASM 磁盘组不包括 Oracle Clusterware LUN，且将装载至 DR 目标站点中已配置的 Oracle Clusterware 堆栈。
- 将 +DATA、+LOG 和 +FRA ASM 磁盘组分开时，可将存储技术用于减轻生产中的备份负担。在热备份流程中，+DATA 和 +FRA 磁盘组将在不同时间克隆。此外，RecoverPoint 和 SRDF 等远程复制将在本机创建数据库可重启的复制副本。可重启的复制副本不会在崩溃或实例恢复期间访问归档日志，因此归档日志（+FRA 磁盘组）无需属于复制。
- +DATA 磁盘组通常可包括临时文件。在测试中使用独立的 ASM 磁盘组进行监视，但这不是特别的部署建议。

表 7. 存储和数据库设备配置和分配

<table>
<tr><th rowspan="2"></th><th colspan="3">精简设备 (LUN)</th><th rowspan="2">数据设备</th></tr>
<tr><th>ASM 磁盘组和 LUN 分配</th><th>精简池绑定</th><th>精简设备</th></tr>
<tr><td>Oracle RAC Grid/ASM 实例</td><td>+Grid ASM 磁盘组</td><td>Redo_Pool</td><td>5 个 20 GB 精简 LUN (15F:163)</td><td rowspan="2">56 个 30GB RAID1</td></tr>
<tr><td rowspan="4">数据库：
名称：
ERPFINDB
大小： 1 TB
Num. LUN： 38</td><td>+REDO ASM 磁盘组</td><td>Redo_Pool</td><td>5 个 20 GB 精简 LUN (164:168)</td></tr>
<tr><td>+DATA： ASM 磁盘组</td><td>Data_Pool</td><td>25 个 60 GB 精简 LUN (1A5:1B4)</td><td>56 个 230 GB RAID5 (3+1) (C5:FC)</td></tr>
<tr><td>+TEMP： ASM 磁盘组</td><td rowspan="4">Temp_Pool</td><td>6 个 50 GB 精简 LUN (17D:182)</td><td rowspan="4">56 个 60 GB RAID5 (3+1) (8D:C4)</td></tr>
<tr><td>+FRA： ASM 磁盘组</td><td>2 个 50 GB 精简 LUN</td></tr>
<tr><td rowspan="2">VPLEX</td><td>VPLEX 元设备</td><td>2 x 2 个 80 GB 精简 LUN (2E5:2E8)</td></tr>
<tr><td>VPLEX 日志</td><td>2 个 50 GB 精简 LUN (2E9:2EA)</td></tr>
</table>

VPLEX Metro 系统的 Symmetrix VMAX 系列存储资源调配

按照以下步骤将存储从 Symmetrix VMAX 系统调配至 VPLEX 虚拟存储环境，这与将存储调配至物理或虚拟服务器的过程基本相同。上述流程假定这是首次将存储从 Symmetrix VMAX 存储调配至 VPLEX Metro 系统。还假定 VPLEX 已分区至 Symmetrix VMAX 存储的前端端口。Symmetrix 操作应该从通过网关守护设备连接至 Symmetrix 的管理主机执行，或使用 Symmetrix Management Console 客户端执行。以下列表

显示的设置活动重点介绍 CLI，但它们也可使用 Symmetrix Management Console 轻松执行。

步骤	操作
1	使用 Symmetrix Management Console 或 Solutions Enabler 命令行界面在本地和远程阵列上创建 Symmetrix 设备。
2	通过执行以下命令，创建 Symmetrix 存储组： **symaccess -sid** *<系统 ID>* **-name** *<组名>* **-type storage devs create** 例如，命令 **symaccess -sid 191 -name VPLEX1_Storage_Group -type storage devs 15F:168 create** 创建存储组 Storage_Group_Test。
3	创建 Symmetrix 端口组。命令： **symaccess -sid 191 -name VPLEX1_Port_Group -type port -dirport** *<目录>:<端口>* **create** 例如，命令 **symaccess -sid 191 -name VPLEX1_Port_Group -type port -dirport 7E:1 create** 创建端口组 Port_Group_Test。
4	创建 Symmetrix 启动器组，其中 VPLEX 后端端口 WWN 是 Symmetrix 启动器组的“主机”启动器。通过运行以下命令，可创建启动器组 Initiator_Group_Test： **symaccess -sid 191 -name VPLEX1_Initiator_Group -type init -wwn** *<WWN>* **create** 例如，命令 **symaccess -sid 191 -name VPLEX1_Initiator_Group -type init -wwn 500014426011ee10 create**
5	创建 Symmetrix 掩蔽视图，对存储、端口和启动器组进行分组： **symaccess -sid 191 create view -name VPLEX1_View -storgrp VPLEX1_Storage_Group -portgrp VPLEX1_Port_Group -initgrp VPLEX1_Initiator_Group**
6	对于要从第二个 VPLEX 系统 (VPLEX2) 调配至第二个 VMAX (sid 219) 的存储，重复步骤 1–5

VPLEX Metro 设置

VPLEX Metro 群集设置步骤

图 8 列出了 VPLEX Metro 设置所需的主要任务。

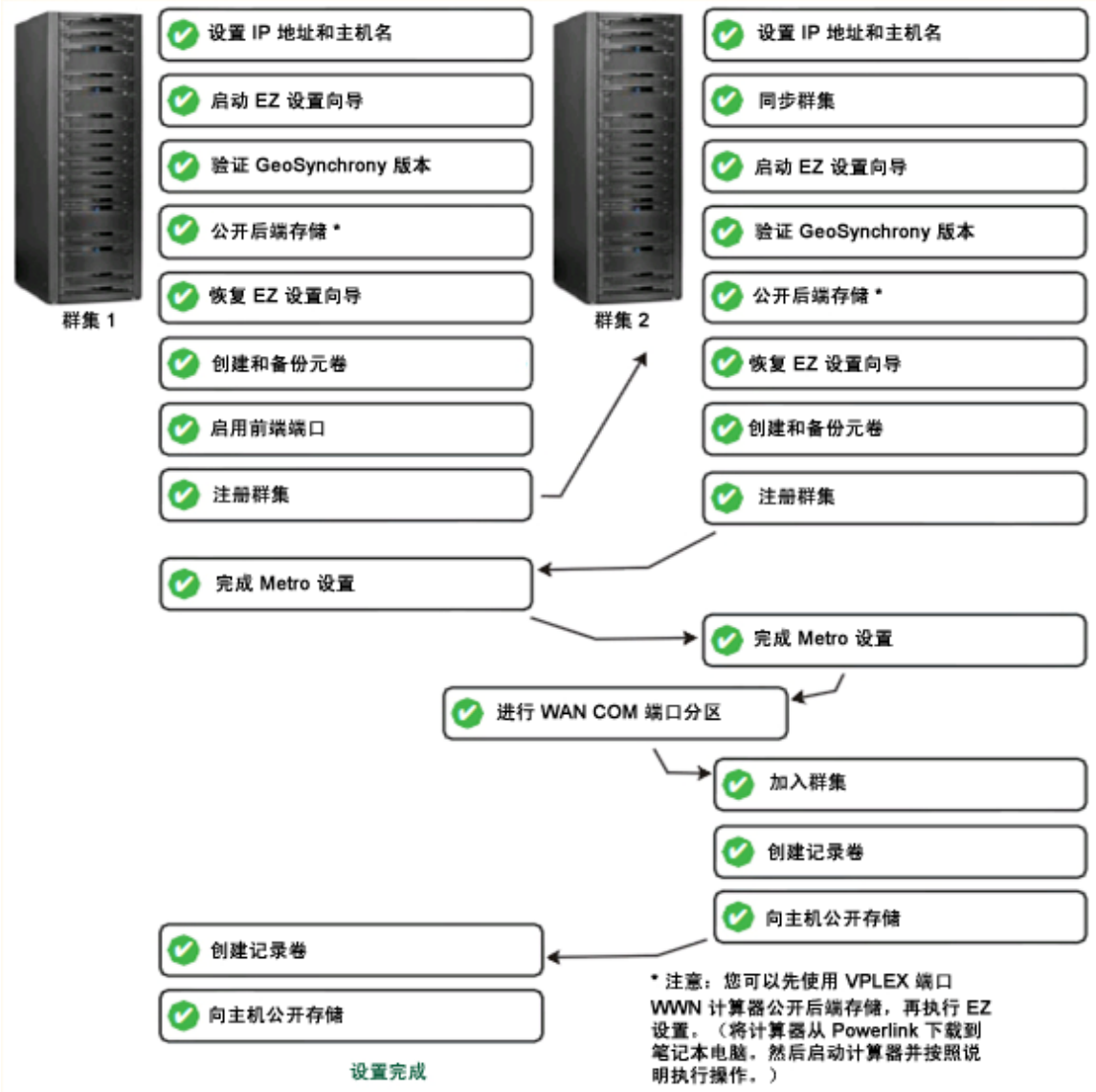


图 8. VPLEX Metro 设置任务概述

注意：您必须按所述方式同时设置两个 VPLEX Metro 群集。您不能单独设置每个群集，再稍后连接它们。

设置 VPLEX Metro 群集连接

VPLEX Metro 站点间通信的两个主要组件是 FC 和 IP。Metro 可以使用光纤通道或 10 Gb 以太网实现每个群集的控制器之间的连接。每个群集的 VPLEX 管理服务器都通过安全 VPN 隧道连接。VPLEX Metro 应该在群集之间设置有冗余（双结构）和完全独立的光纤通道或 10 Gb 以太网网络，以进行控制器间通信。这可以提供最高的性能、故障隔离、容错能力及可用性。图 9 是群集间 WAN 连接的分区示例。

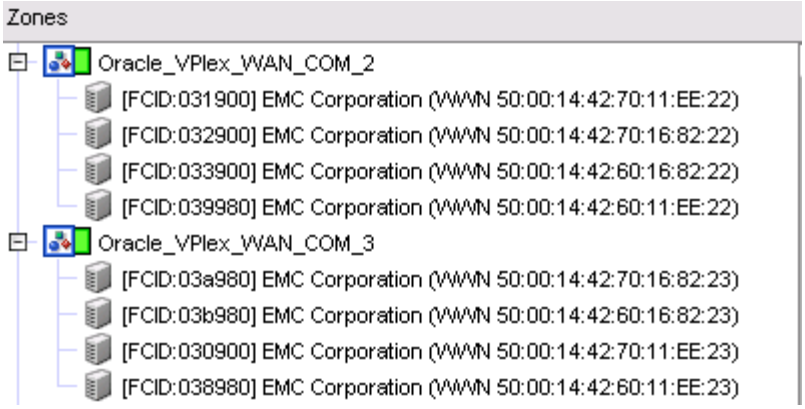


图 9. VPLEX 群集间 WAN 连接的分区示例

检查群集连接

要检查 WAN 连接，请登录 VPLEX CLI 并运行以下命令：

ll **/hardware/ports/

示例：

```
VPlexcli:/> ll **/hardware/ports/
/engines/engine-1-1/directors/director-1-1-A/hardware/ports:
Name     Address              Role       Port Status
-------  ------------------   ---------  -----------
A2-FC00  0x500014426011ee20   wan-com    up
A2-FC01  0x500014426011ee21   wan-com    up
A2-FC02  0x500014426011ee22   wan-com    up
A2-FC03  0x500014426011ee23   wan-com    up

/engines/engine-1-1/directors/director-1-1-B/hardware/ports:
Name     Address              Role       Port Status
-------  ------------------   ---------  -----------

B2-FC00  0x500014427011ee20   wan-com    up
B2-FC01  0x500014427011ee21   wan-com    up
B2-FC02  0x500014427011ee22   wan-com    up
```

```
B2-FC03  0x500014427011ee23  wan-com    up

/engines/engine-2-1/directors/director-2-1-A/hardware/ports:
Name     Address             Role       Port Status
-------  ------------------  ---------  -----------
A2-FC00  0x5000144260168220  wan-com    up
A2-FC01  0x5000144260168221  wan-com    up
A2-FC02  0x5000144260168222  wan-com    up
A2-FC03  0x5000144260168223  wan-com    up

/engines/engine-2-1/directors/director-2-1-B/hardware/ports:
Name     Address             Role       Port Status
-------  ------------------  ---------  -----------
B2-FC00  0x5000144270168220  wan-com    up
B2-FC01  0x5000144270168221  wan-com    up
B2-FC02  0x5000144270168222  wan-com    up
B2-FC03  0x5000144270168223  wan-com    up
```

要检查 FC MAN 链路状态，请运行 **cluster summary** 命令。

示例：

```
VPlexcli:/> cluster summary
Clusters:
Name       Cluster ID  Connected  Expelled  Operational Status Health State
---------  ----------  ---------  --------  ------------------ ------------
cluster-1  1           true       false     ok                  ok
cluster-2  2           true       false     ok                  ok
Islands:
  Island ID  Clusters
  ---------  --------------------
  1          cluster-1, cluster-2
```

VPLEX Metro 主机连接

为了确保扩展 Oracle RAC 的最高级别连接性和可用性（甚至是将 Oracle RAC 服务器连接至 EMC VPLEX 的异常操作期间），扩展 Oracle RAC 部署模式中的每台 Oracle RAC 服务器都应该至少具有两个物理 HBA，而且每个 HBA 都应该连接至 EMC VPLEX 不同控制器上的前端端口。此配置可确保即使在 EMC VPLEX 的一个前端端口因计划内维护事件或计划外中断而离线时，也能继续使用 Oracle RAC 节点。

将单 VPLEX 引擎配置连接到扩展 Oracle RAC 节点时，每个 HBA 都应同时连接至 VPLEX 引擎内 A 和 B 控制器上提供的前端端口。到 VPLEX 前端端口的连接应包括：先将具有唯一性的主机连接到模拟前端控制器的每个 I/O 模块的端口 0，然后将其

他主机连接到 I/O 模块的其余端口。如果多个 VPLEX 引擎可用，Oracle RAC 服务器中的 HBA 应该连接至不同引擎。

从 VPLEX 引擎到存储阵列的连接应该遵守针对阵列的最佳做法建议。关于连接后端存储的最佳做法的详细讨论超出了本白皮书的范围。《EMC VPLEX Architecture and Deployment: Enabling the Journey to the Private Cloud》（EMC VPLEX 体系结构和部署：支持私有云之旅）技术书籍提供了详细信息。

VPLEX Metro 管理

运行 Geosynchrony 5.1 的 VPLEX Metro 可通过 EMC Unisphere™ for VPLEX Management Console 进行管理。某些附加高级功能通过 VPLEX CLI 提供。对基于 Web 的安全 GUI 进行身份认证时，将向用户提供一组按完成顺序列示的屏幕配置选项。EMC Unisphere for VPLEX Management Console 联机帮助提供了有关工作流中每个步骤的详细信息。下表汇总了从发现阵列到存储对主机可见这个过程中要执行的步骤。

步骤	操作
1	**发现可用存储** VPLEX Metro 自动发现连接至后端端口的存储阵列。所有连接至群集中每个控制器的阵列都会列示在存储阵列视图中。
2	**申请存储卷** 必须先申请存储卷，然后才能将其用于群集（但是元数据卷除外，这种卷通过未申请的存储卷来创建）。仅当申请存储卷后才能将其用于创建扩展区、设备和虚拟卷。
3	**创建扩展区** 为选定存储卷创建扩展区并指定容量。
4	**从扩展区创建设备** 简单设备从一个扩展区创建且仅使用一个群集中的存储。
5	**创建虚拟卷** 使用上一步中创建的设备创建虚拟卷。
6	**注册启动器** 直接或通过光纤通道结构连接启动器（访问存储的主机）时，VPLEX Metro 会自动发现它们并填充**启动器视图**。必须先向 VPLEX Metro 注册所发现的启动器，然后才能将其添加至存储视图和访问存储。注册启动器将对端口的 WWN 指定有意义的名称（通常是服务器的 DNS 名称），让您可以轻松标识主机。
7	**创建存储视图** 要让存储对主机可见，请先创建存储视图，然后将 VPLEX Metro 前端端口和虚拟卷添加至视图。虚拟卷要等到填入具有关联端口和启动器的存储视图之后才可见。

8	**创建一致性组** 在两个 VPLEX 系统上创建一致性组，并且将所有为扩展 Oracle RAC ASM 设备分配的虚拟卷添加至一致性组，包括 Oracle Database 中要求写入顺序一致性的 Grid 和 ASM 设备。

图 10 显示关于逻辑布局以及从 EMC VPLEX 调配存储的联机帮助。

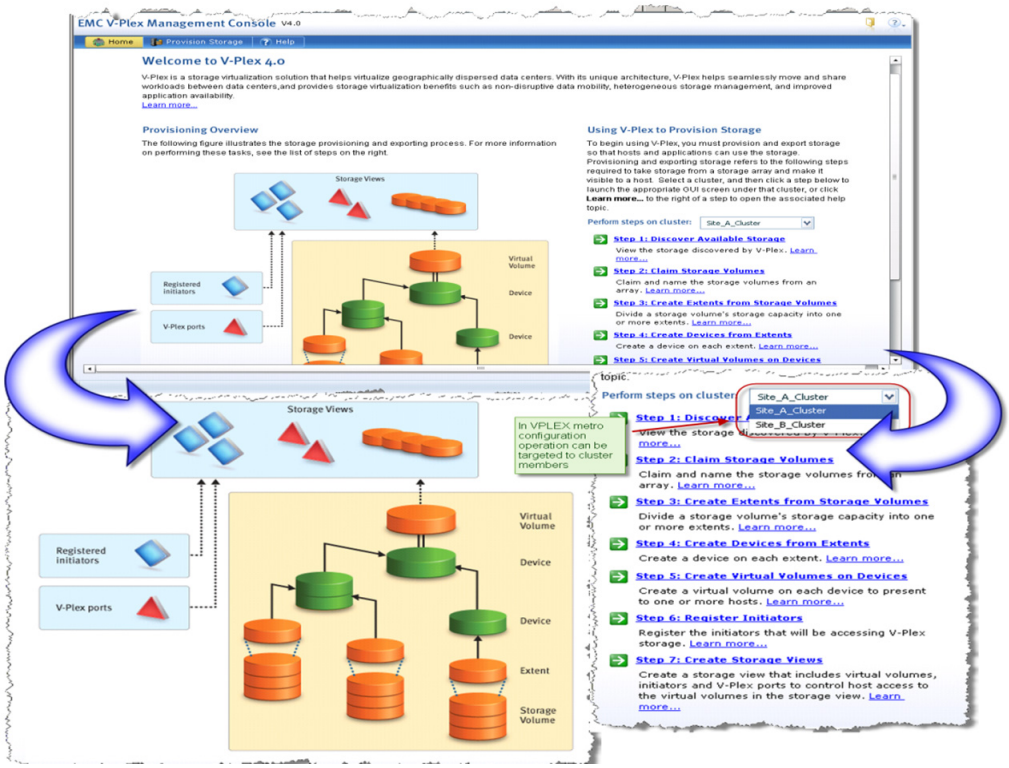


图 10. EMC Unisphere for VPLEX 管理界面

基于浏览器的管理界面用图解展示了此过程中涉及的各个组件，如图 10 所示。EMC VPLEX 中的存储通过一种称为“存储视图”的逻辑结构公开；存储视图是“注册的启动器”、“VPLEX 端口”和“虚拟卷”这三种对象的联合体。“注册的启动器”对象列出了需要访问存储的启动器的 WWPN（全球通用端口名称）。若是 Oracle VM 服务器环境，“注册的启动器”实体包含 Oracle VM 服务器中连接到 EMC VPLEX 的 HBA 的 WWPN。“VPLEX 端口”对象包含 VPLEX 阵列的前端端口，“注册启动器”通过这些端口访问虚拟卷。“虚拟卷”对象是使用由后端存储阵列提供给 EMC VPLEX 的存储卷构成的卷的集合。从图 10 左下角的插图中可以看到，虚拟卷由“设备”构成，而后者是由基于一种抽象实体（称为“扩展区”）构建的不同设备组合而成。该图还显示“扩展区”是从向 EMC VPLEX 公开的“存储卷”中创建的。但是，为了利用基于阵列的复制技术，我们将每个存储设备从其整个（一对一映射）直通配置中的 VMAX（设备容量 = 扩展区容量 = 存储卷容量）映射到

VPLEX，并且采用 RAID 0（仅限单个扩展区）VPLEX 设备结构。这样，VPLEX 便不会影响底层存储设备，而且后端阵列 LUN 复制技术（如 TimeFinder/Clone 和 TimeFinder/Snap）继续正常运行。

图 10 右下角的插图中还显示了从 EMC VPLEX 中调配存储所需的七个步骤。该向导支持一种集中化调配机制，可在 EMC VPLEX Metro 环境中向不同的群集成员调配存储。从 EMC VPLEX 中调配存储的第一步是发现与其连接的存储阵列。很少需要执行这一步，因为 EMC VPLEX 会主动监视存储环境的变化。此过程的第二步是“声明”向 EMC VPLEX 公开的存储。存储申请过程将创建如图 10 所示的对象“存储卷”。“创建存储视图”向导让您可以创建存储视图，并将启动器、端口和虚拟卷添加到视图。将所有组件都添加到视图后，它将自动变为活动状态。在存储视图处于活动状态时，主机可看到存储并开始对虚拟卷的 I/O。创建存储视图后，只能通过 GUI 添加或删除虚拟卷。要添加或删除端口和启动器，请使用 CLI。EMC VPLEX CLI 指南提供了有关 VPLEX Metro 命令的综合信息。

带 VPLEX Witness 的 VPLEX Metro

VPLEX Witness 作为一个封闭的虚拟机安装，且部署在与 VPLEX 群集不同的故障域中（以消除单个故障同时影响群集和 VPLEX Witness 的可能性）。VPLEX Witness 通过管理 IP 网络连接至两个 VPLEX 群集。VPLEX Witness 通过将其自身的观察与群集定期报告的信息进行协调，让群集可区分群集内网络分区故障和群集故障，并在这些情况下自动恢复 I/O。

主机和 Oracle 设置

多路径软件设置

在将 Oracle RAC 服务器连接至 VPLEX 的最佳做法中，每台 Oracle RAC 服务器都应该具有两个 HBA 端口，并且每个端口都连接至不同 FC 交换机，以提高可用性。在此类配置中，主机必须使用多路径解决方案来处理到同一存储设备的多条路径，以提供高可用性、负载平衡和实时迁移。您可以为 Oracle RAC 服务器安装 EMC PowerPath 作为多路径解决方案，也可以使用 Linux 本机多路径解决方案（设备映射程序）。对于 EMC VPLEX Metro 配置中的扩展 Oracle RAC（如本白皮书所述），EMC PowerPath 5.5 安装在四台物理服务器上。

在每个主机上安装 PowerPath rpm

```
[root@ RAC NODE 1: licoc039 ]rpm -i EMCpower.LINUX-5.5.0.00.00-
275.RHEL5.x86_64.rpm
```

首次为主机安装 PowerPath 之后，可能需要重启才能注册 /dev/emcpower 伪设备。

在每个主机上安装 PowerPath 许可证

```
[root@ RAC NODE 1: licoc039 ]emcpreg -add <密钥>
```

在每个主机上配置 PowerPath

```
[root@ RAC NODE 1: licoc039 ] powermt config
[root@ RAC NODE 1: licoe039 ] powermt display
…
Pseudo name=emcpowerk
Invista ID=FNM00100600231
Logical device ID=6000144000000010A002636D3C679C6A
state=alive; policy=ADaptive; priority=0; queued-IOs=0
==============================================================================
---------------- Host ----------------   - Stor -   -- I/O Path -  -- Stats ---
### HW Path                I/O Paths    Interf.   Mode    State  Q-IOs Errors
==============================================================================
   1 lpfc                      sdaq      08        active  alive      0      0
   2 lpfc                      sdbu      00        active  alive      0      0
   2 lpfc                      sdcy      08        active  alive      0      0
   1 lpfc                      sdm       00        active  alive      0      0
```

跨 Oracle RAC 服务器节点匹配 PowerPath 伪设备名称

要跨 Oracle RAC 服务器节点匹配 PowerPath 伪设备名称，EMC 建议使用 PowerPath 应用工具 **emcpadm**。此应用工具可用于从一个主机导出映射并将其导入另一个主机。如有必要，它还允许一次重命名一个伪设备。

```
<源主机> emcpadm export_mapping -f <映射文件名>
```

接着将该文件拷贝到其他主机。使用存储设备关闭任何应用程序，卸载任何文件系统，或者导出任何 LVM 卷，然后运行：

```
<目标主机> emcpadm check_mapping [-v] -f <映射文件名>
<目标主机> emcpadm import_mapping -f <映射文件名>
```

在 PowerPath 设备上创建分区

EMC 强烈建议在使用 Symmetrix 时，以 64 KB 偏移量对齐基于 x86 的服务器平台分区。使用其他存储阵列时，尽管大小应该总是以 4 KB 偏移量对齐以匹配 VPLEX 数据块大小，但是可能具有不同要求。图 11显示了以 64 KB 对齐以在 PowerPath 设备上创建分区，如下所示简单启动 fdisk，并且在创建分区后，键入“x”进入专家模式。键入“p”显示（打印）当前分区表，包括以数据块单元表示的偏移量。键入“b”更改任何分区偏移量。例如，将分区 1 从其默认偏移量 32 个数据块移至 128。由于每个数据块都是 512 字节，因此 128 x 512 字节 = 64 KB 偏移量。如果在 LUN 上创建多个分区，请确定其余分区已对齐，或者按照类似步骤将其偏移量更改为以 128 个数据块 (64 KB) 对齐的数字。

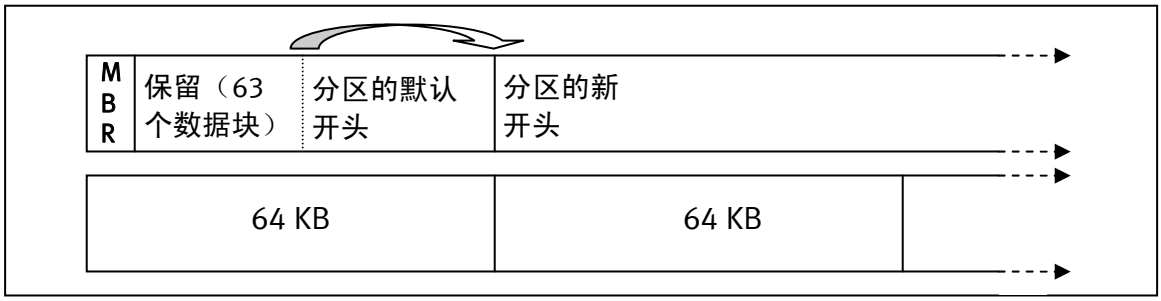


图 11. Symmetrix 跟踪大小边界的分区对齐 (64 KB)

在本示例中，为将要用作 Oracle ASM 设备的 PowerPath 设备创建了一个分区。

```
[root@licoc091 ~]# fdisk /dev/emcpowerd
Command (m for help): n
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-52218, default 1): [ENTER]
Using default value 1
Last cylinder or +size or +sizeM or +sizeK (1-52218, default 52218):
[ENTER]

Command (m for help): p

Disk /dev/emcpowerd: 54.7 GB, 54755328000 bytes
64 heads, 32 sectors/track, 52218 cylinders
Units = cylinders of 2048 * 512 = 1048576 bytes

   Device           Boot    Start       End      Blocks    Id  System
/dev/emcpowerd1            1       52218     53471168    83  Linux
```

以 64 KB 边界（128 个数据块）对齐分区。

```
Command (m for help): x（进入专家模式）

Expert command (m for help): p（打印分区表）
请注意，分区 1 从 32 个数据块开始

Disk /dev/emcpowerk: 64 heads, 32 sectors, 52218 cylinders

Nr AF  Hd Sec  Cyl  Hd Sec  Cyl     Start      Size ID
 1 00   1   1    0  63  32 1023        32    106942336 83
 2 00   0   0    0   0   0    0         0         0 00
 3 00   0   0    0   0   0    0         0         0 00
 4 00   0   0    0   0   0    0         0         0 00
Expert command (m for help): b（移动分区开头）
Partition number (1-4): 1
New beginning of data (32-2002943, default 32): 128

Expert command (m for help): p

Disk /dev/emcpowerk: 64 heads, 32 sectors, 52218 cylinders
```

```
Nr AF  Hd Sec  Cyl  Hd Sec  Cyl      Start       Size ID
 1 00   1   1    0  63  32 1023        128  106942336 83
 2 00   0   0    0   0   0    0          0          0 00
 3 00   0   0    0   0   0    0          0          0 00
 4 00   0   0    0   0   0    0          0          0 00

Expert command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.
[root@licoc091 ~]#
```

创建分区后，确保其他节点可识别它们。可能需要在其他每个节点上运行 fdisk 命令并写入（“w”）分区表。或者，重新扫描 SCSI 总线或重启其他节点也将刷新信息。

安装 Oracle 并设置 Oracle RAC 数据库

下表汇总了根据《Oracle 11g Release 2 (11.2.0.2) Grid Infrastructure Installation Guide for Linux》（适用于 Linux 的 Oracle 11g 版本 2 (11.2.0.2) Grid Infrastructure 安装指南）和《11g Release 2 (11.2.0.2) Database Installation Guide for Linux》（适用于 Linux 的 11g 版本 2 (11.2.0.2) 数据库安装指南），为 Oracle Grid Infrastructure 和 ASM 数据库安装配置 Oracle 服务器节点所需的步骤。Oracle RAC 安装指南提供更多详细信息。详细的 Oracle 安装指南可从此处获得：

http://www.oracle.com/technetwork/documentation/index.html#database

主要步骤：

步骤	操作
1	配置服务器节点专用网络、OS /etc/hosts 文件、OS 内核参数，然后编辑 /etc/sysctl.conf 文件。
2	创建用于在每个 RAC 节点上安装和维护 Oracle 11gR2 的 Oracle 用户组和帐户。
3	更新启动脚本 (/etc/rc.d/rc.local) 以设置对 Oracle Clusterware 和 Oracle ASM 指定的设备的 Oracle 权限。
4	在每个 Oracle RAC 节点上为 Oracle 用户设置 ssh。
5	在 /etc/security/limits.conf 文件中为 Oracle 用户设置 shell 限制。
6	相应修改 /etc/pam.d/login 文件和 /etc/profile 文件。
7	安装 Oracle 安装所需的任何附加 OS 软件包。
8	安装 Oracle 11g 版本 2 (11.2.0.2) 数据库软件。
9	为 Oracle Database 创建附加 Oracle ASM 磁盘组。
10	使用满足所需数据库工作负载和性能要求所需的大小和初始化参数创建 Oracle 数据库。为 Oracle Database 创建附加 Oracle ASM 磁盘组。

OLTP 数据库工作负载测试

我们使用标准 Oracle OLTP 工作负载（分别为 70/30 随机读/写比率）演示了 VPLEX Metro 群集在 Metro 距离达到 100 km 的位置向扩展 Oracle RAC 提供高性能和工作负载平衡能力。如本白皮书的实施部分所述，采用 VPLEX Metro 的扩展 Oracle RAC 测试环境包含两个本地 Oracle RAC 节点和两个远程 Oracle RAC 节点。VPLEX 群集和 Oracle RAC 节点之间的 WAN 互连是使用 Empirix PacketSphere Network Emulator 实现网络延迟的模拟 Metro 距离（高达 5ms RTT）。如图 12 和图 13 所示，16 个驱动器工作负载是从每个 Oracle RAC 节点执行的，而且针对单个节点、两个节点、三个节点或四个节点工作负载记录了 OLTP 工作负载事务率（每分钟事务数）。每添加一个节点都会增加工作负载，这展示了 Oracle RAC 的可扩展性好处。随着添加的 Oracle RAC 节点越来越多，事务率也随着工作负载的增加而成比例增大（1-4 个节点分别为 16、32、48 和 64 个工作负载驱动器）。在距离为 0 km 和 Metro 距离达到 100 km 的两种情况下，采用 Oracle RAC 的 VPLEX Metro 显示了类似的事务率增长。此外，VPLEX Metro 能够提供高性能，其中扩展 Oracle RAC 在 Metro 距离达到 100 km 时的事务率约为 0 km 距离时的 85-90%。此外，如果 Metro 距离在 50 km 以内，VPLEX Metro 将能针对扩展 Oracle RAC 实现超过 90% 的工作负载性能（未显示数据）。因此，VPLEX Metro 可针对两个数据中心的 Metro 距离为 100 km 的 OLTP 工作负载，向 Oracle RAC 提供高 I/O 性能和恢复能力。测试表明，500 km 距离 (5 ms RTT) 也能显示类似的可扩展性，但是，事务率会因延迟增加而相对较低。总之，该解决方案证明了 VPLEX 和 Oracle RAC 都能够提高应用程序性能和可用性。请注意，由于 OLTP 基准测试完全随机，因此未遇到数据块争用情况。在具有实际客户工作负载的部署中，DBA 应该关注（总是如此）群集节点之间可能降低总体事务率的潜在数据块争用情况，尤其是远程节点之间发生的争用。

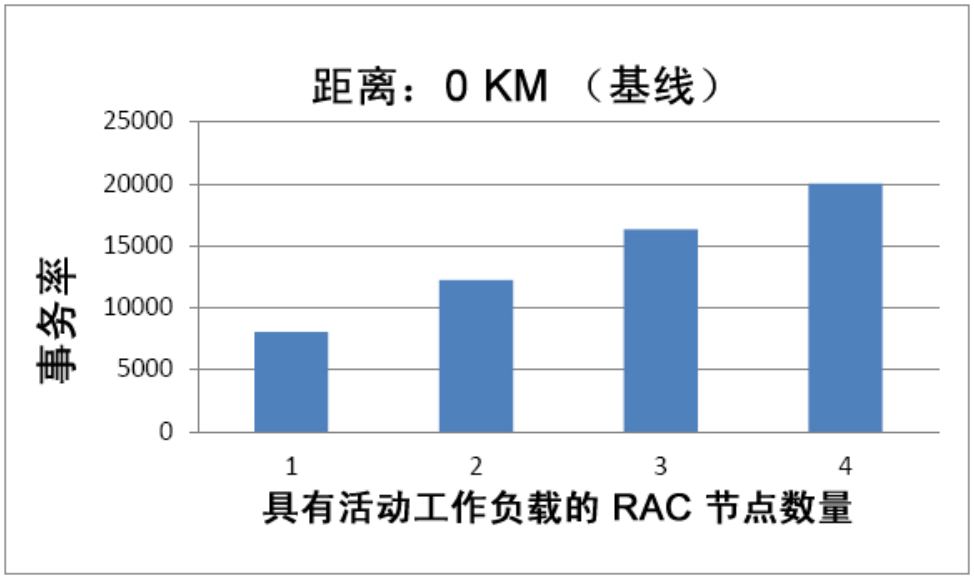


图 12. 0 km 距离时的事务率

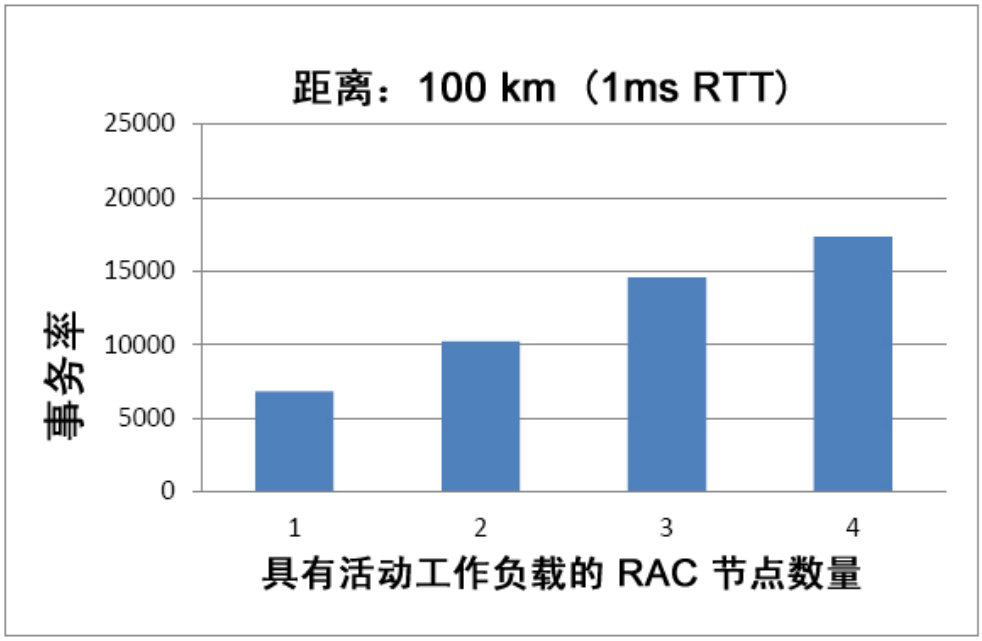


图 13. 100 km 距离时的事务率

故障情形测试

对于项目团队 E-Lab™（EMC 质量和资格认证组织）驱动的众多故障情形，环境都成功通过了基于 Oracle RAC 工程部门提供的测试计划的测试。所有测试都成功完成，且达到预期结果。所有测试都是在运行事务工作负载时执行的（适用时）。执行的一部分测试列表包括：

- VPLEX Metro 基础架构保持不变时的 Oracle RAC 互连分区（“Split-Brain”）。
- Oracle RAC 互连和 VPLEX Metro 互连分区。
- 单个存储系统断开连接，但没有应用程序宕机。
- 仍正常运行的站点继续运行工作负载的站点故障模拟。
- 不同 Oracle RAC 节点的主机连接中断，但不影响群集其余部分。
- Oracle ASM 重新平衡、存储以及 VPLEX 配置更改和软件更新。

结论

运行 EMC GeoSynchrony 操作系统的 EMC VPLEX Metro 是一种基于 SAN 的企业级联合技术，可聚合和管理光纤通道连接的存储阵列池，这些阵列可共存于同一个数据中心，也可分布于地理位置相隔 Metro 距离的多个数据中心。而且，凭借独特的纵向扩展和横向扩展体系结构，EMC VPLEX 高级数据缓存和分布式缓存吻合性提供了存储域的工作负载恢复能力、自动共享、平衡和故障切换，并通过可预测服务级别支持本地和远程数据访问。如果扩展 Oracle RAC 分散于 Metro 距离内的两个数据中

心（由 EMC VPLEX Metro 功能提供支持），则可提供简化的部署拓扑和存储管理、无中断的存储可扩展性和技术刷新。另外，EMC VPLEX 还可在同步距离内提供无中断的异构数据移动和卷管理功能，让客户能够跨越多个物理位置提供灵活、经济且高效的云服务。

参考资料

以下文档包括有关 VPLEX 和扩展 Oracle RAC 的详细信息，并且可在 http://china.emc.com、Docs.Oracle.com 和 Powerlink 上找到：

- *EMC VPLEX 实施和规划最佳做法技术说明*
- 《EMC VPLEX Metro Witness — 技术和高可用性》技术书籍
- EMC VPLEX Metro 上的扩展主机群集支持条件
- 《EMC VPLEX with GeoSynchrony 5.0 and 5.1 Product Guides》（带 GeoSynchrony 5.0 和 5.1 的 EMC VPLEX 产品指南）
- http://www.oracle.com/goto/rac
- *Oracle Database 11g 版本 2 (11.2)*
- *Oracle Database 11g Interactive Quick Reference*